Academic Year 2011/2012

UNIVERSITY OF NAPLES "FEDERICO II"

PhD In Computational biology and Bioinformatics

XXIV CICLO

Genome-complexity in terms of gene amplification: intriguing issues from reference species

Coordinator:

Prof. Sergio Cocozza

Student:

Alessandra Vigilante

Tutor:

Dott.ssa Maria Luisa Chiusano

Co-tutor:

Prof.ssa Paola Festa

Abstract

Gene duplication (GD) and alternative splicing (AS) have emerged as two major processes supporting the functional diversification of the genes.

My research project is mainly focused on the genome-scale analysis of these two important and complex mechanisms. Using comparative approaches and bioinformatics strategies I focused on the analysis of intragenome duplications, specifically focusing on Transcription Factors gene families in Plants and *Homo sapiens*. In particular, to investigate on gene duplication in Plants, the *A.thaliana* genome was used as a reference. The *Arabidopsis* genome underwent ancient whole genome duplication events (WGDs), followed by gene reduction and diploidization [Blanc et al., 2000; Vision et al., 2000; Simillion et al., 2002; Cui et al., 2006; reviewed in Van de Peer and Meyer, 2005]. However, what dramatically increases its complexity are the extended genome rearrangements (i.e. deletions, inversions, translocations), which relocated and split up the retained portions around the genome [Tang et al., 2008], together with probable chromosome reductions within the Brassicaceae [Conner et al., 1998].

Under the classical model for the evolution the duplicated genes may go for loss of function, neofunctionalization and subfunctionalization. In the first case, one member of the duplicated pair usually degenerates by accumulating deleterious mutations, while the other copy retains the original function. In the case of neofunctionalization, one duplicate may acquire a new adaptive function and the result is the preservation of both members of the pair, one with presenting the new function and the other retaining the old one. Functional divergence can occur even by subfunctionalization, that is the two copies act with a complementary effect to accomplish the functionalities of the ancestral gene. Duplicated genes also may interact through inter-locus recombination, gene conversion, or concerted evolution.

In particular, we designed a bioinformatics pipeline to detect duplicated and singleton genes, taking into account several issues related to the computational methods applied. The implemented pipeline can provide a reference as tool for the detection of paralogy relationships in other genomes. Moreover, set of genes sharing one or more paralogs were organized in networks made available to the scientific community for small and large scale analyses, while single copy genes were deeply investigated since their presence represents an intriguing aspect in a so highly duplicated genome.

A web accessible database (available at http://biosrv.cab.unina.it/athparalogs/main/index) allows access to the network organization, and the relevance of this resource either for evolutionary investigation or gene family analyses is here presented. In addition, our analysis underlines the need of a more accurate annotation process for the *Arabidopsis* genome and stirs up intriguing evolutionary issues related to the presence of single copy genes in a highly duplicated genome.

Since the release 10 of the *Arabidopsis* genome [The Arabidopsis Information Resource, 2010] was recently made public, we confirmed and briefly described the main results concerning the TAIR9 also for this newest version. Transcription factor gene families (TFs) were analyzed considering the collection of networks obtained from *A. thaliana*. Due to their key roles in gene regulation, TFs are among the best examples of dosage-sensitive genes.

This work provides support to the classification of transcription factors in *A.thaliana* and represents a step forward to understand TF families organization and evolution. Transcription factors were also analyzed in terms of alternative splicing. To further investigate on the impact of alternative splicing on transcriptional regulation, a genome-wide study of alternative splicing of TFs was also considered in the human genome, providing insights into the dynamic usage of splice isoforms and their regulatory impact in different cell types.

Content

Chapter 1

Introduction

1.1 *Arabidopsis thaliana* as a reference plant species

Evolutionary and comparative genetics between plant species have been supported by the use of reference species as models for the purpose of understanding plant biology. The complete genome sequences and gene–trait associations revealed for the references species have provided enormous insight into all plant species, their chromosomes, genes, pathways, evolution and hence relationships to one another, providing a preliminary framework for understanding the genetic and molecular diversity in plants and plant processes. It is only a beginning because of the immense diversity across the plant kingdom, and, Because of this diversity, the concept of one or a few species being reference "models" for all species is limited. Moreover, since an ideal model should be endowed of relevant information achieved more quickly and cheaply than other species (1, 2), some of key features of an initial model should be valid. The main features are shown in Table 1.

Table 1. Preferred attributes of a model plant species
Attributes
Small genome
Rapid life cycle
Easily transformed
Diploid genetics with few chromosome/gene duplication
Well positioned in plant phylogeny
Small stature for growth in small space
Large number of seeds produced
Convenient for discovery of gene-trait linkages at low cost, high speed

Table 1 Preferred attributes of a model plant species.

The table shows a list of preferential features of model plant species

These are the reasons why in 1980 *Arabidopsis thaliana* became the leading contender around the world (3– 6) after some debate about *Petunia* and some other species. Friedrich Laibach had studied *Arabidopsis* from the early 1900s, and Erna Rheinholz in the early 1940s, but it was Glass (7), Redei (8) and Koornneef (9) who opened up mutational genetics in the species. *Arabidopsis*, classified within the eudicots lineage of flowering plants, inevitably has major limitations as both a model and a framework reference for monocots, that occurs in the other major lineage of flowering plants (Figure 1). That is why rice plays such an important role for understanding monocots and their genomes, complementing *Arabidopsis* for the study of angiosperms in general. The success of *Arabidopsis* as the leading model species and its value can be inferred from the number of publications and the databases devoted to the species since 1985.

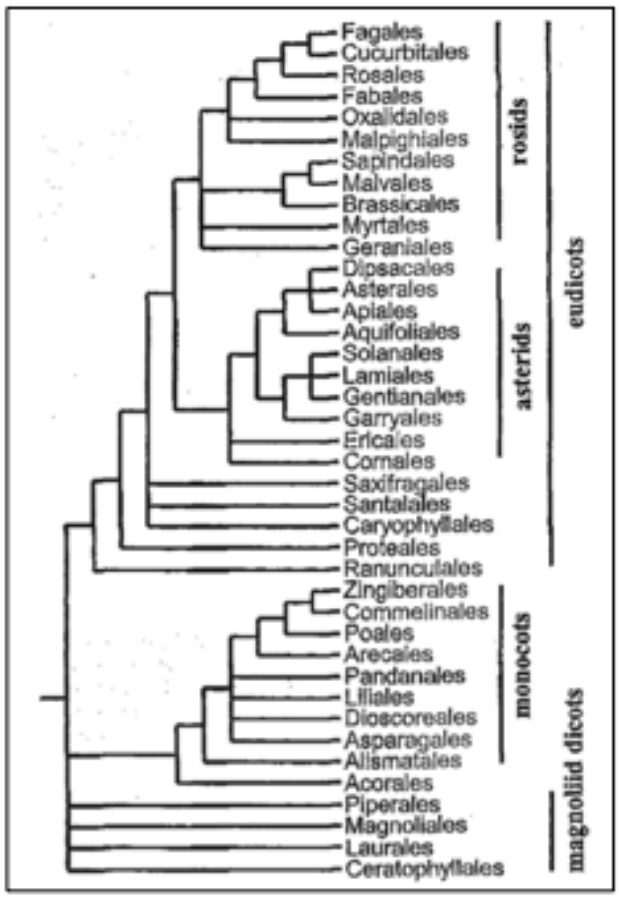


Figure 1 Angiosperm phylogeny.

Arabidopsis is in *Brassicales* of the rosids, and rice is in *Poaces* of the monocots. Modified from Angiosperm Phylogeny Group. Adapted from Somers [Somers et al., 2009].

There have always been and will be arguments about the suitability of one species as a model for others. Protein sequences are relatively highly conserved across species but their coding sequences are frequently reused with variant promoters and other regulatory sequences to provide functional diversity. While it is expected that closely related proteins will carry out the same function in different species it is obviously expected that mutations in coding sequences within and between species will diversify basic functions somewhat by changing affinities for substrates, binding affinities to other proteins, metabolites, DNA and RNA complexes, etc. To discover this, homologs, paralogs and potential orthologs can be screened similarly within the model species to understand the extent to which diversity in coding sequences has led to differences in function. Thus, a comprehensive understanding of the model genome is mandatory.

The problem is when genomes diverge over time, they accumulate mutations that include not only base changes in specific genes, but also changes in the number and distribution of repeated sequences, including transposable elements. Such changes create huge numbers of chromosomal differences within and between species resulting in major changes in DNA content, but not necessarily in the order of genes along chromosomal segments. Though, plant breeding depends on the frequency of recombination between genes and so knowing the order of genes in linkage blocks is very useful. Gene order is conserved during evolution and thus reflects the phylogenetic relationships between species. Indeed, knowing the order of genes along a chromosomes segment of the model species can be a guide to the order of genes along the evolutionary equivalent chromosomal segment of another related species.
In the light of this evidence, a reference genome ideally should be reliable, well understood in terms of organization and evolutionary history and well annotated. Apparently, this is not the case of *A.thaliana*: it's genome is really gene dense, not yet exhaustively annotated and moreover it resulted a highly complex genome since it underwent at least three rounds of whole genome duplications [], followed by reduction and reshuffling of its gene content. In this context, the unraveling of the *Arabidopsis thaliana* genome organization in terms of duplicated regions is one of the major challenges of the plant genomics as the fulfillment of this aims is valuable for understanding it's evolutionary history and ii) making a full use of *Arabidopsis thaliana* as reference plant species in comparative genomics.

To provide a contribution to these challenges, this thesis is mainly focused on a genome-scale analysis of the *Arabidopsis thaliana* genome in terms of duplicated and single copy genes. This chapter first gives a general introduction to the process of gene and gene duplication as mechanisms, which lead to gene amplification and protein diversity. Different duplication mechanisms will be described as well as the possible fate of duplicated genes. Then the chapter introduces alternative splicing as another important mechanism for proteomic diversity. The third section gives an overview of eukaryotic transcription factor genes. According to the "gene balance hypothesis" these genes, which are key regulators of the transcriptional machinery, are considered dosage sensitive genes, i.e. genes encoding proteins that are needed in stoichiometric amounts, as part of a multi-protein complex. Dosage sensitivity is an important evolutionary force, which impacts on gene dispensability and duplicability.

In this frame it is extremely interesting to understand how this class of genes is affected by such mechanisms which drive protein diversification. The chapter concludes with a summary of the aims of this thesis.

1.2. Mechanisms to achieve protein diversification

Understanding how protein structures and functions have diversified is one of the central goal in molecular evolution. Extensive research has established that gene duplication is intimately connected with the creation of new genes and with the innovation of protein function during evolution. Ohno [][][] was a strong supporter of the idea that redundant genes are invaluable templates for innovating protein functions. He suggested that there were at least two whole genome duplications (WGDs) during vertebrate evolution and that duplicate genes provide raw genetic material for novel functions. This proposal was extended by Kimura and Ota [], who insisted "gene duplication must always precede the emergence of a gene having a new function".

Gene sharing is another important source of new protein functions. In general, the term "gene sharing" was established to describe the fact that one gene produces a polypeptide that has more than one molecular function: in other words two or more entirely different functions of a protein share the identical gene. The gene-sharing concept postulates that protein function is determined not only by primary amino acid sequence, which remains the same in the multiple functions that are performed by the protein, but also by the microenvironment within the cell and by the expression of its genes. As a process, gene sharing "tests" for new protein functions. A gene may be engaged in gene sharing with its encoded polypeptide performing the same multifunctional tasks for eons of time. Gene sharing almost certainly serves as one of the mechanisms for retaining duplicated genes, along with subfunctionalization and Darwin selection []. Indeed, the acquisition of new functions by gene sharing operates independently of gene duplication and may occur in single-copy genes, duplicate genes, or individual members of gene families (Figure 2).

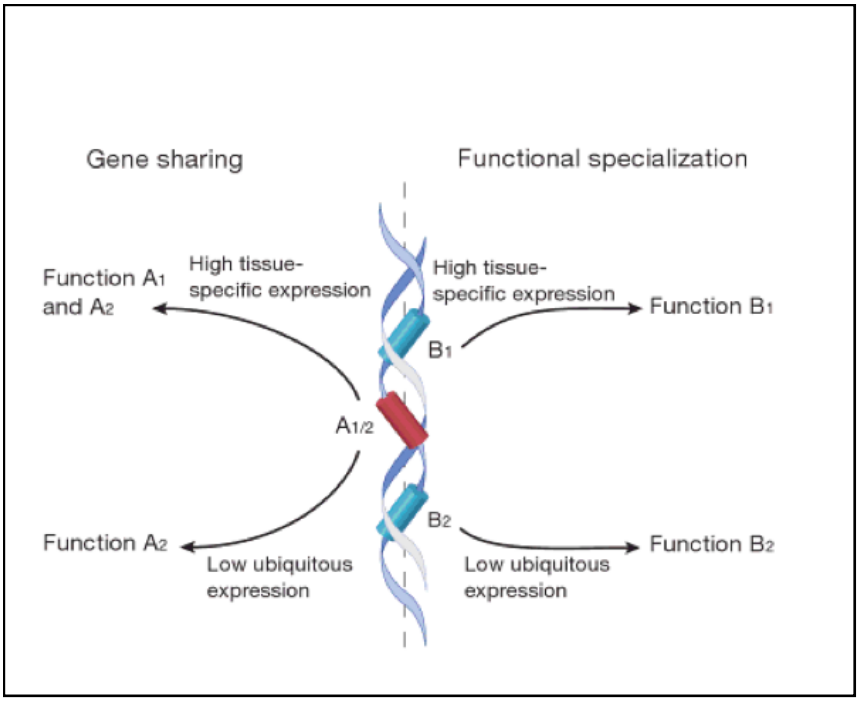


Figure 2 Gene sharing versus gene duplication followed by functional specialization.

The left side of the figure illustrates one gene (A) encoding a protein with the ability to perform two functions (A1 and A2). The protein performs functions A1 and A2 when gene A is highly expressed in a specific tissue, but it only works as A2 when expressed at a low level. The right side of the figure illustrates functional specialization of daughter genes (B1 and B2) after gene duplication. In this case, functions B1 and B2 are performed by distinct proteins encoded by duplicated genes. Adapted from Piatigorsky [Piatigorsky et al., 2009].

In addition to the fundamental biological processes of gene/genome duplication and gene sharing, many eukaryotic organisms have evolved other means to achieve protein diversity.

One of the most common mechanisms for generating diversity by a single gene is alternative RNA splicing of its primary transcript. Alternative splicing generates two or more distinct mRNAs by different elimination of introns when processing the primary transcript into the mature mRNA. Different mRNA sequences can therefore be translated by complex biochemical processes into proteins, which may be different in terms of amino acid sequences. The resulting proteins can have different expression patterns and variable biological roles (Figure 3).

Another mechanism generating protein diversity involves initiating transcription at alternate sites from a stretch of DNA. This is called alternative promoter utilization because the promoter of a gene lies in front of and attached to the sequence where

transcription is initiated. In Figure 3, the mechanisms is represented by having either promoter 1 initiate transcription at the beginning of exon 1 or promoter 2 (located within intron 1) initiate transcription at exon 2.

Although infrequent, RNA editing is another mechanism that affects gene function. Conversion of one nucleotide into another in the protein-coding sequence of the mRNA itself by an editing process, results in a different amino acid sequence after translation of the edited mRNA (Figure 3).

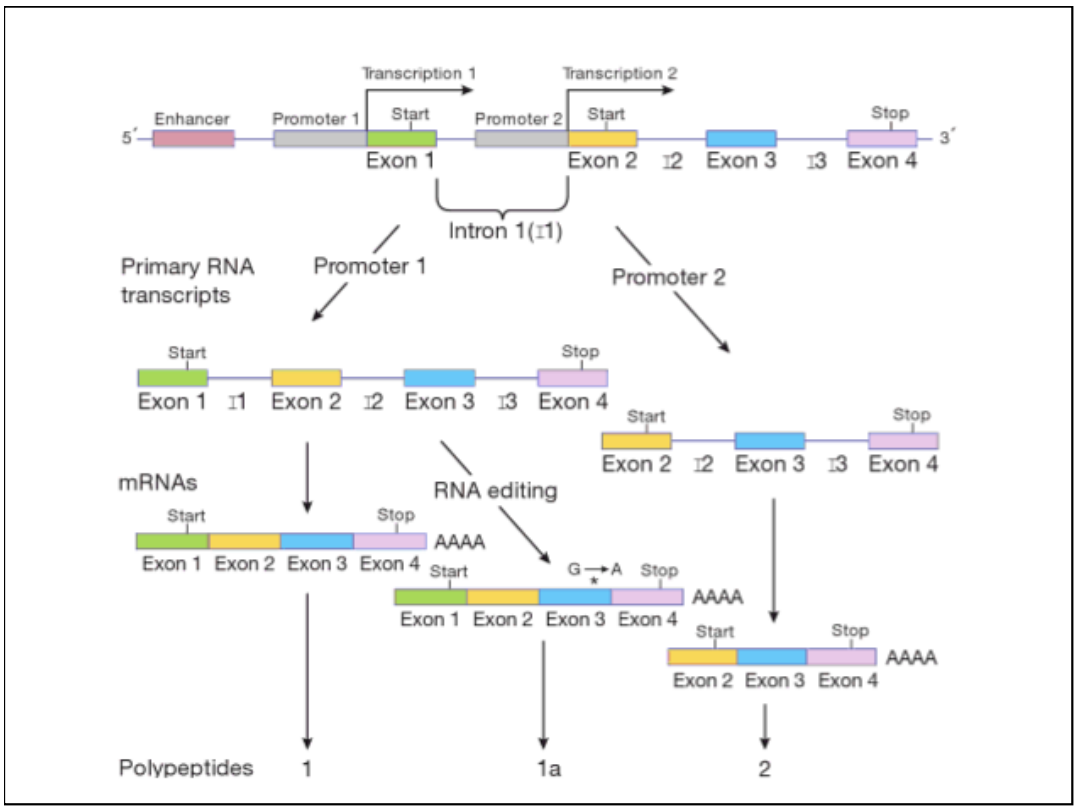


Figure 3 Protein diversity from a single gene.

Three proteins (1, 1a, 2) with different amino acid sequences are produced from one gene by alternative RNA splicing and RNA editing. Adapted from Piatigorsky [Piatigorsky et al., 2009].

Among all the events listed above, alternative splicing and gene duplication are largely considered as the two main contributors to the diversity of the protein repertoire with enormous impact on protein sequence, structure, and function [1–5].

1.2.1 Gene duplication and protein evolution

Gene and genome duplications have been thought to play an import role in evolution since the 1930s (Bridges 1936; Stephens 1951; Ohno 1970). Recent genomic sequence data provide substantial evidence for the abundance of duplicated genes in many organisms.

Duplication events produce additional copies of genomic information, usually involving one or more genes. While in some cases these duplicated elements may be of immediate benefit (i.e. increasing availability and effective dosage of a desired gene product), often they are, initially at least, somewhat redundant, and either neutral or mildly detrimental to the fitness of the organism. It is of no surprise, then, that the majority of duplicated genes are quickly deactivated by mutations abolishing their transcription or translation. Some duplicated genes, however, survive and persist, suggesting that their retention has indeed some benefit. Many of these genes seem to have acquired properties that distinguish them from their progenitors; they may be expressed in a novel tissue type, for example, or differ in their functional specificity. In these cases, it appears as though duplication has facilitated evolution, either by allowing specialization and refinement or, perhaps most intriguingly, generating genes free to mutate and acquire 'novel' functions. These retained duplicates form a family of genes related through common ancestry. As a result of their common origin, gene sequences within gene families are often quite similar, complicating the task of assigning them unique and specific functions. As a consequence, there has been a significant effort to study and characterize the evolution of function in the consequences of a duplication event. The following sub-paragraphs will briefly overview the various modes of gene duplication, and then will focus on the various functional outcomes of duplication.

1.2.2 Mechanisms of duplication

There are several different mutational mechanisms through which gene duplicates can be produced. Depending on the type of event, the nature and scale of what is duplicated can differ significantly. Single genes may be copied, with or without their peripheral regulatory elements, as well as entire genomes or chromosomes can be duplicated . While each mechanism ultimately results in the duplication of one or more genes, the mechanisms differ in three key respects: how much regulatory

information the duplicated genes retain; where the duplicates are integrated into the genome; how many interaction partners are duplicated. Duplication mechanisms can be broadly categorized into three groups – DNA/RNA-mediated transposition, unequal recombination, and genome/chromosome doubling (whole genome/chromosome duplication). All these mechanisms produce paralogs, i.e. homologous genes that are both present in and native to the same genome (in contrast to orthologs, where speciation acts as a 'duplication event' and the homologous genes are components of different genomes). Figure 4 provides a diagram depicting various modes of duplication.

1.2.2.1 DNA/RNA transposition

DNA/RNA transposition refers to mechanisms by which a specific short nucleotide sequence, either mRNA (as in retrotransposition) or DNA (e.g., transposon-mediated duplication) is copied from one location in the genome to another. The insertion is essentially random. RNA-mediated retrotransposition is unique in that it uses post-transcribed sequences as a template for the nascent duplicate. Hence, upstream and downstream regulatory sequences lying outside the transcribed gene sequence are not preserved, and the newly produced copy will reflect the structure of the mature RNA, with most or all introns (and possibly some exons) spliced out. The new gene may also possess a genetically encoded poly-A tail. Since RNA-mediated retrotransposition does not preserve most non-coding regulatory elements, the duplicate gene must depend on the a-priori availability or acquisition of promoter/regulatory sequences in order to be transcribed. Absence of these elements effectively means that the new gene duplicate is, at least initially, a pseudogene. DNA-mediated duplications, as mediated by transposons, for example, often retain regulatory information and intron/exon structure. Nonetheless, they still operate on a very specific subsequence of DNA, and elements relocated by DNA-mediated transposition can be inserted in any eligible location in the genome.

1.2.2.2 Segmental duplication/unequal cross-over

Errors during homologous recombination can produce serial duplications of genetic

sequence. Unequal crossing-over is an error stemming from the miss-alignment of homologous chromosomes during mitosis/meiosis. Ordinarily, homologous sequences are aligned and cross-over events result in balanced exchanges of sequence information across chromosomes. An abundance of repetitive sequences can, however, cause chromosomes to misalign, in which cases a segment of one chromosome is inserted into its sister chromatid (thus producing a duplication and a reciprocal deletion). Since multiple rounds of unequal crossing over tend to gradually inflate the number of candidate repeat regions, some genomic regions are hotbeds for sequence duplication and can give rise to a large number of duplicate genes in series. These serially arranged duplicates are referred to as "tandem duplicates". These tandem gene arrays are highly evident in the genome, and tandem duplicates retain most or all of their intron/exon structure and peripheral non-coding elements. Unequal crossing-over also plays a role in the generation of copy number variations (Redon et al., 2006).

1.2.2.3 Whole genome/chromosome duplications

In some circumstances, errors during segregation can produce diploid gametes, and the fusion of these diploid gametes can result in a complete doubling of genomic content (all chromosomes present in duplicate). While very rare, these whole genome duplication (WGD) events have a dramatic impact on the content of the genome. Different numbers of WGD events have been hypothesized in the history of various lineages (Van de Peer et al., 2009) in both plant and animals, and they had interesting implications for the evolution of gene regulation (Lockton & Gaut, 2005). By their nature, WGD events result in the duplication of all loci, preserving non-coding elements, intron/exon structure, and even overall stoichiometry within gene/protein interaction networks. Interestingly, it has been observed that lineages that underwent separate, distinct WGD events often ultimately retain similar (i.e. orthologous) duplicates – that is to say, WGD duplicates that becomes fixed in one lineage, were also often fixed in the other (Semon & Wolfe, 2008).

WGD events are relatively common in plant lineages. Allopolyploids are a variant of whole genome duplications in which the diploid gametes come from two different species. These genomic hybrids contain two formerly independent complete genomes.

The most commonly studied allopolyploids are plants, though a number of examples have been documented elsewhere in the animal kingdom (including the model organism Xenopus Laevis). Duplicates produced through allopolyploidy (i.e. formerly orthologous genes now present in the same organism) are often referred to as "homeologs" (Flagel et al., 2008).

In general, polyploidy is widespread in plants (Vision et al., 2000; Adams & Wendel, 2005), presumably aided by their ability to propagate vegetatively and by the existence of specific regulatory mechanisms in plant cells. Estimates of the incidence of polyploidy in angiosperms vary from 30 to 80%, and about 3% of speciation events are explained by genome duplications (Otto & Whitton, 2000). Many, if not all, species of plants may thus have at least one polyploid ancestor. Most eudicots are assumed to have an ancient hexaploid ancestor, with subsequent tetraploidization in some taxa (Jaillon et al., 2007). Epigenetic silencing may protect the duplicated copies from pseudogenization, thus facilitating the acquisition of new functions (Rodin & Riggs, 2003). In particular, model polyploid plants have been characterized by a rapid loss of some genes and the specific inactivation of others by methylation (Kashkush et al., 2002; Comai et al., 2000; Lee & Chen, 2001).

Finally, chromosomal duplications can take place. In this case extra copies of a chromosomal region are formed, resulting in different copy numbers of genes within that area of the chromosome. If the duplicated sections are adjacent to the original, the process is known as tandem duplication, whereas if they are separated by non duplicated regions, the duplication is said to be displaced.

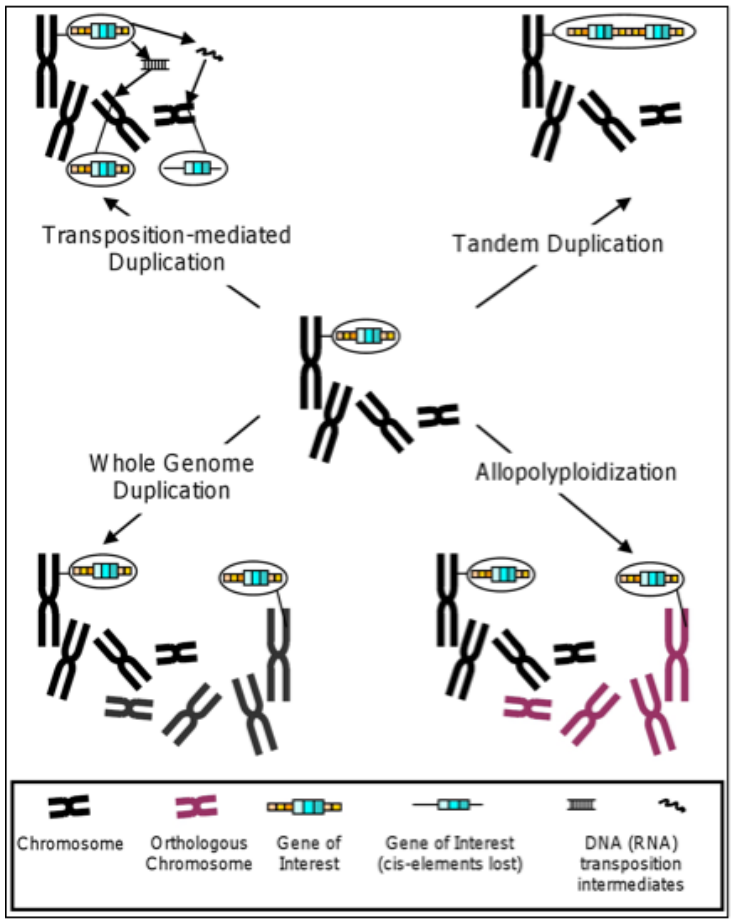


Figure 4 Modes of gene duplication.

Upper left: transposition mediated by either a DNA or RNA intermediate can produce gene at distant locations in the genome. RNA intermediates retain little of the regulatory sequence surrounding the parent gene. Upper right: Errors during homologous recombination can produce tandem arrays of genes, situated in series. Bottom Left: Doubling of all chromosomes will produce duplicates of all genes in the genome. Bottom Right: Allopolyploids contain genomes from two compatible species, with duplicate gene pairs from formerly orthologous. Adapted from Friedberg [Friedberg et al., 2011]

1.2.3 Theoretical models for duplicate retention and functional specialization

A number of theoretical models have been proposed to describe how a parental gene function can be partitioned between offspring, and how this partitioning affects the chances of these genes to avoid pseudogenization and eventual deletion. Three

different fates outcomes, specifically: nonfunctonalization, subfunctionalization, and neofunctionalization, which are based on concepts typically attributed to Ohno (1970). Figure 5 depicts two hypothetical duplications and their respective functional specializations.

1.2.3.1 Nonfunctionalization

Nonfunctionalization describes the situation where the expression of one duplicate is abolished, making it invisible to natural selection and thus free to accumulate mutations. While it is technically possible for a nonfunctionalized gene to have its function restored, the vast majority become relics progressively crippled by the accumulation of disabling and deleterious mutations. There has been some interest in studying the impact losing a duplicate via nonfunctionalization has on sibling genes – for example, whole genome duplication events can lead to cases of "orthologs gone missing", where a WGD duplicate has been lost (Canestro et al., 2009). Reciprocal duplicate loss has been hypothesized as one mean of speciation.

1.2.3.2 Neofunctionalization

Neofunctionalization refers to the scenario where one duplicate gene acquires mutations that allow it to acquire previously unexplored functions, either through changes in regulation (e.g. tissue localization) or in coding potential. Claims of neofunctionalization tend to focus on the generation of new functions, though it should be noted that these developments might also result in the loss of ancestral function(s) (Turunen et al., 2009). A specific example of neofunctionalization can be found in a recent study of the MADS-box gene family in angiosperms. MADS-box genes are well known for their role in developmental processes, but the functions of some gene family members have been difficult to determine. Viaene et al. (2010) provide evidence that a group of these genes, the AGL6 subfamily, can be neatly divided into two groups based on duplication history. One of these groups retains the ancestral function of guiding reproductive development, while the other seems to have acquired a novel role in regulating the growth of vegetative tissues.

1.2.3.3 Subfunctionalization

Subfunctionalization involves each gene taking upon a complementary subset of the parental gene functionalities, such that each copy is not independently capable of fulfilling all the parental gene's roles. Subfunctionalization is conceptually synonymous with the Duplication, Degeneration, and Complementation (DDC) model. Regulatory subfunctionalization could result in non-overlapping tissue distributions for the nascent duplicates, with the union of the expression profiles matching the parental gene's range. Jarinova et al. (2008) describe an instance of subfunctionalization of the Hox genes of zebrafish. Through a careful analysis of peripheral non-coding elements, the authors show how the two hoxb complexes in zebrafish, hoxb5a and hoxb5b, acquired non-overlapping expression profiles. In particular, the experimental removal of one regulatory element unique to hoxb5a resulted in the two paralogs (re)acquiring a similar expression profile.

The idea of structural subfunctionalization is perhaps best captured in the "Escape from Adaptive Conflict" (EAC) hypothesis. Consider a hypothetical gene product with multiple interaction partners (e.g. an enzyme with two possible substrates). Selection for bi-functionality in this enzyme may limit the binding/catalytic efficiency of either specific reactions -- mutations that improve one may inhibit the other, hence the "adaptive conflict". Should this gene be duplicated, however, each offspring gene could be free to acquire mutations that optimize binding to one specific substrate, thus escaping the conflict without a loss of functionality. The EAC model essentially describes this process, where a single enzyme with multiple interaction partners gives rise to duplicate genes with more specific but enhanced functionality.

EAC is interesting in that it lies somewhere on the boundary between subfunctionalization and neofunctionalization. Three claims are required to invoke the model: that i) both duplicates accumulate adaptive changes post mutation, that ii) these mutations enhance ancestral functions, and lastly that iii) the ancestral gene was constrained from improving functions (Barkman & Zhang, 2009). The key difference (and challenge) lies in proving the ancestral form was bi-functional. Studies demonstrating the EAC model in action are still relatively uncommon. An early attempt to apply the model to the genes from the anthocyanin biosynthetic pathway of morning glories has come under criticism for not clearly providing these three veins of supporting evidence (Barkman & Zhang, 2009; Des Marais & Rausher, 2008).

While duplicated genes are generally relegated to one of the fates listed above, a number of case studies have shown that recent duplicates can maintain identical functional profiles. One possible explanation for this is that the duplicates have acquired mutations that have restored the "status quo" that was present prior to duplication. If mutations cause the sum of the duplicate genes' expression levels to be equal to the expression level of their ancestor, both genes could experience some level of selective pressure to maintain expression despite being fully redundant. Ganko et al. (2007) observed that a vast majority of duplicates, regardless of duplication mechanisms, showed asymmetric expression, with one gene consistently showing higher levels of expression, This suggests that a limited form of subfunctionalization may play an initial role in the retention of duplicates.

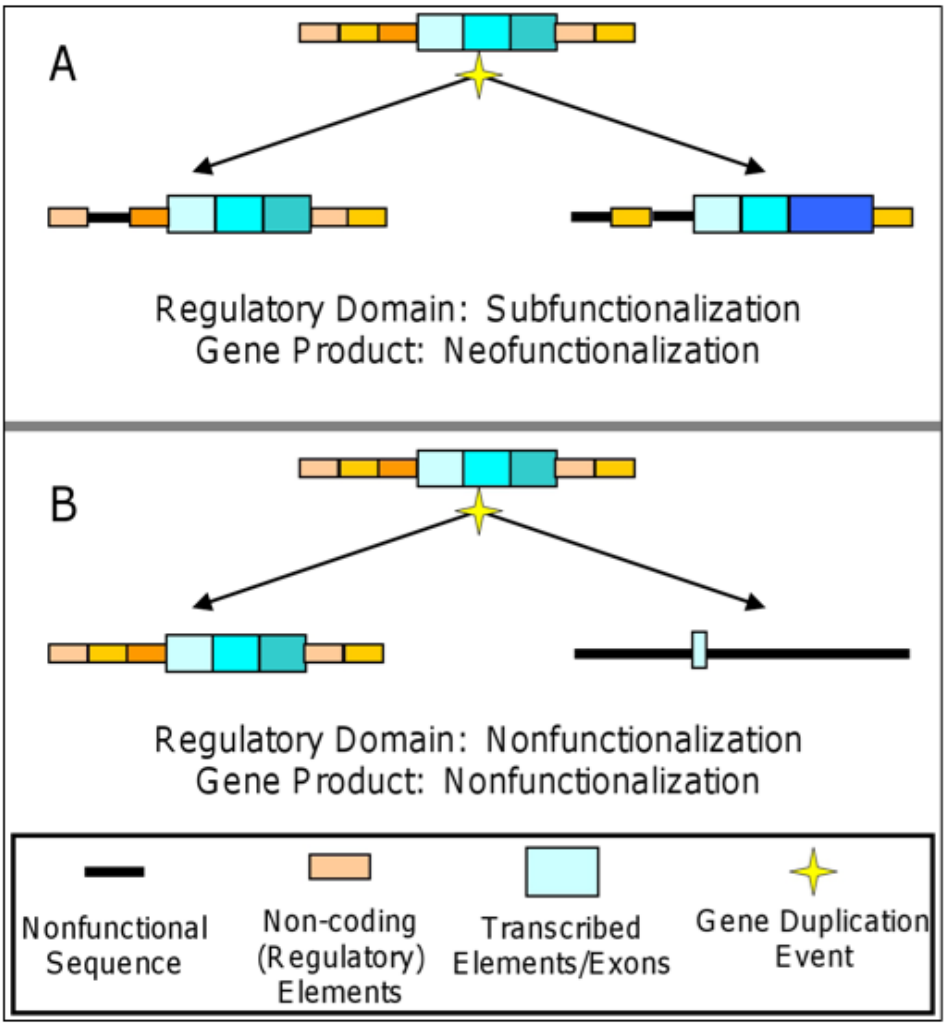


Figure 5. Possible functional specializations following duplication.

Two hypothetical examples showing how retention models can apply either to regulatory regions or gene products. **A)** Duplicated genes subfunctionalize at the regulatory level, partitioning their parental regulatory domains and suggesting subdivided roles. The gene product, however, has acquired a novel element (i.e. new exon), suggesting neofunctionalization at the coding sequence level. **B)** Following duplication, one gene loses its regulatory domains and is interrupted by an early stop codon, reflecting nonfunctionalization both at the regulatory and gene product levels. Adapted from Friedberg [Friedberg et al., 2011].

1.2.3.4 Reductions and rearrangements after WGDs: increase in complexity

Detection of natural ancestral polyploidy is a difficult task, especially for very ancient events. Recent duplications can be detected by comparing closely related species, one of which underwent diploidization and therefore contains twice as many chromosomes as species that did not undergo WGD. The older is the duplication, the harder is the analysis, because a period of diploidization often follows polyploidization, which "transforms" the polyploid genome to the diploid state. Diploidization is achieved by an intensive loss of genes, rearrangements of the genome and divergence of duplicated genes. Recent analyses have also shown that the duplication of individual genes in evolution has occurred much more frequently than was previously thought (Lynch & Conery, 2000; Lynch et al., 2001). Diploidization has been studied in many genomes including those of yeasts (Piskur, 2001; Kellis et al., 2004; Scannell et al., 2006; Scannell et al., 2007), Paramecium (Aury et al., 2006), vertebrata (Blomme et al., 2006) and plants (Chapman et al., 2006; Jaillon et al., 2007; Tuskan et al., 2006). In particular, among the latter *Arabidopsis* provide a clear example: since the earlier analyses after the completion of *Arabidopsis* genome, revealed a huge number of rearrangements of its gene content, resulting in a patchwork of duplicated regions that indicated deletion, insertion, tandem duplication, inversion and reciprocal translocation. The grapevine/*Arabidopsis* comparison also confirmed that the *Arabidopsis* genome lineage has undergone many rearrangements and chromosome fusions such that the ancestral gene order is particularly difficult to deduce from this species (Figure 6).

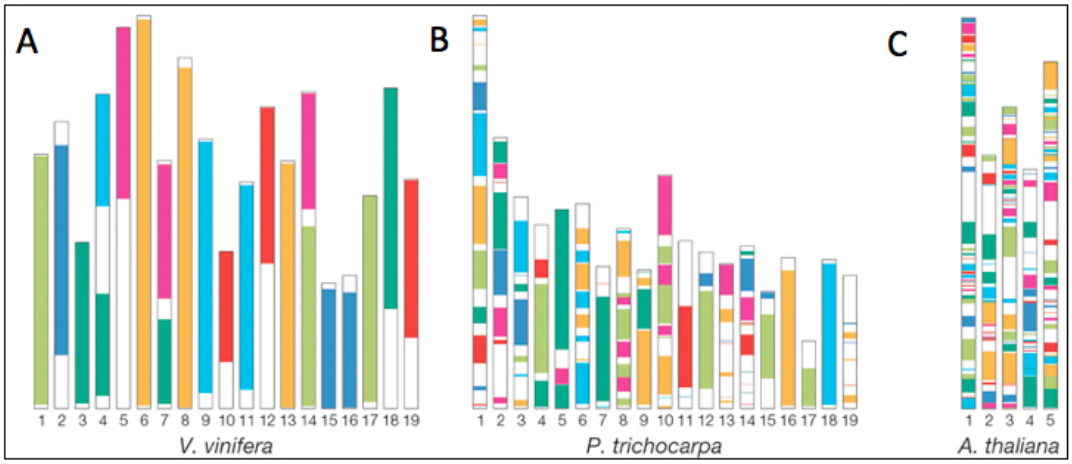


Figure 6. Schematic representation of paralogs regions derived from the three ancestral genomes in the karyotypes of V. vinifera, P. trichocarpa and A. *thaliana*. Adapted from Jaillon [Jaillon et al., 2007].

Each color corresponds to a syntenic region between the three ancestral genomes that were defined by their occurrence as linked clusters in grapevine, independently of intrachromosomal rearrangements. The *V. vinifera* genome (**A**) is by far the closest to the ancestral arrangement, whereas that of *Arabidopsis* (**C**) is thoroughly rearranged, and *P. trichocarpa* (**B**) presents an intermediate situation. The seven colors probably correspond to linkage groups at the time of the palaeo-hexaploid ancestor.

1.3 Alternative splicing: another mechanism to increase complexity

Prior to genomics, studies of alternative splicing primarily focused on the mechanisms of alternative splicing (AS) in individual genes and exons. This has changed dramatically since the late 1990s. High-throughput genomics technologies, such as EST sequencing and microarray designed to detect changes in splicing, and more recently the deep sequencing provided by next generation technologies, led to genome-wide discoveries and quantification of alternative splicing in a wide range of species from human to *Arabidopsis*. Consensus estimates of AS frequency in the human genome grew from less than 5% in mid-1990s to as high as 60-74% today. The rapid growth in sequence and microarray data for alternative splicing has made it

possible to look into the global impact of AS on protein function and evolution of genomes. The next subparagraphs will briefly describe the alternative splicing mechanisms, and then will focus on their impact on proteomic complexity and their role in genome evolution.

1.3.1 Insights into the mechanism of alternative splicing

Alternative splicing is a process that allows the production of a variety of different proteins from one gene only (3, 4) (Figure 7). Most genes in eukaryotic genomes consist of exons and introns. After transcription, introns need to be removed from the pre-mRNA by a step called splicing. Sometimes an exon can be either included or excluded from the final transcripts, or there can be two splice sites at one end of an exon that are recognized by the spliceosome (the complex which determines the splicing reaction). In a typical multiexon mRNA, the splicing pattern can be altered in many ways (Figure 7). Most exons are constitutive; they are always spliced or included in the final mRNA. A regulated exon, that is sometimes included and sometimes excluded from the mRNA, is called a cassette exon. In certain cases, multiple cassette exons are mutually exclusive, as they produce mRNAs that always include one of several possible exon choices but no more than these. In these systems, special mechanisms must enforce the exclusive choice [][]. Exons can also be lengthened or shortened by altering the position of one of their splice sites. Both alternative 5′ and alternative 3′ splice sites are possible. The 5′-terminal exons of an mRNA can be switched through the use of alternative promoters and alternative splicing. Similarly, the 3′-terminal exons can be switched by combining alternative splicing with alternative polyadenylation sites. Alternative promoters are primarily an issue of transcriptional control. Control of polyadenylation appears mechanistically similar to control of splicing, although this is not further discussed here []. Finally, some important regulatory events are controlled by the failure to remove an intron, a splicing pattern called intron retention. Particular pre-mRNAs often have multiple positions of alternative splicing, giving rise to a family of related proteins from a single gene (Figure 7H).

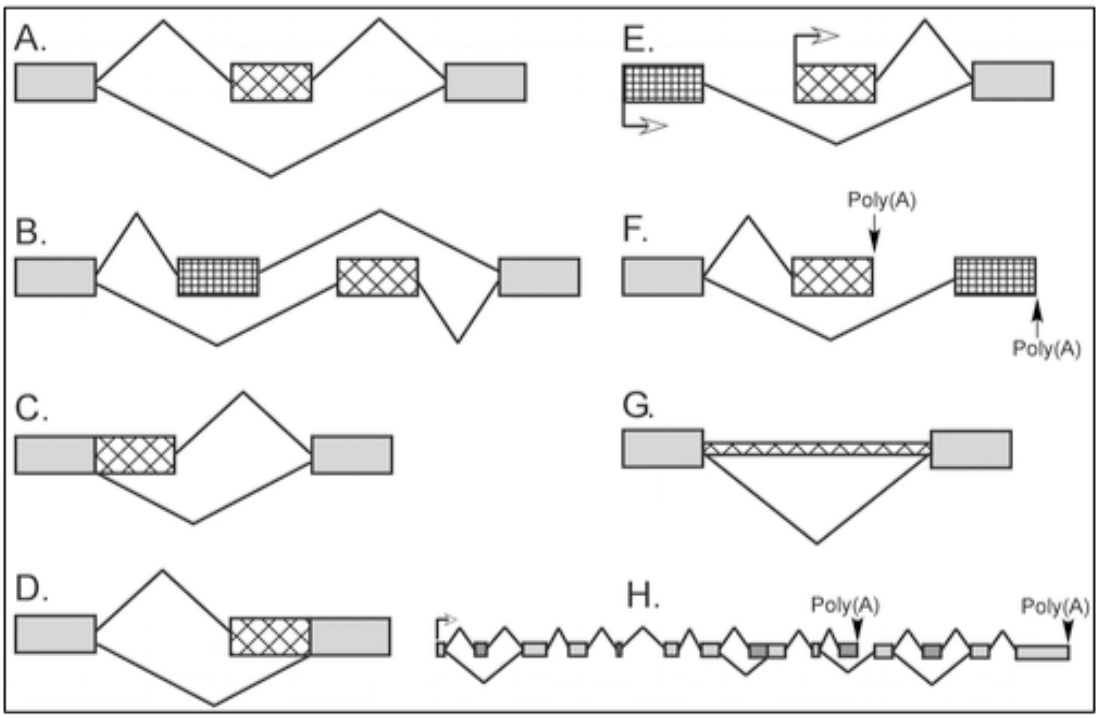


Figure 7. Patterns of alternative splicing.

Constitutive sequences present in all final mRNAs are gray boxes. Alternative RNA segments that may or may not be included in the mRNA are hatched boxes. **A.** A cassette exon can be either included in the mRNA or excluded. **B.** Mutually exclusive exons occur when two or more adjacent cassette exons are spliced such that only one exon in the group is included at a time. **C, D.** Alternative 5′ and 3′ splice sites allow the lengthening or shortening of a particular exon. **E, F.** Alternative promoters and alternative poly(A) sites switch the 5′- or 3′-most exons of a transcript. **G.** A retained intron can be excised from the pre-mRNA or can be retained in the translated mRNA. **H.** A single pre-mRNA can exhibit multiple sites of alternative splicing using different patterns of inclusion. These are often used in a combinatorial manner to produce many different final mRNAs.

The excision of the introns from a pre-mRNA and the joining of the exons is directed by special sequences at the intron/exon junctions called splice sites []. The 5′ splice site marks the exon/intron junction at the 5′ end of the intron (Figure 8A). This includes a GU dinucleotide at the intron end encompassed within a larger, less conserved consensus sequence []. At the other end of the intron, the 3′ splice site region has three conserved sequence elements: the branch point, followed by a

polypyrimidine tract, followed by a terminal AG at the extreme 3′ end of the intron. Splicing is carried out by the spliceosome, a large macromolecular complex that assembles onto these sequences and catalyzes the two transesterification steps of the splicing reaction (Figure 8). In the first step, the 2′-hydroxyl group of a special A residue at the branch point attacks the phosphate at the 5′ splice site. This leads to cleavage of the 5′ exon from the intron and the concerted ligation of the intron 5′ end to the branch-point 2′-hydroxyl. This step produces two reaction intermediates, a detached 5′ exon and an intron/3′-exon fragment in a lariat configuration containing a branched A nucleotide at the branch point. The second transesterification step is the attack on the phosphate at the 3′ end of the intron by the 3′-hydroxyl of the detached exon. This ligates the two exons and releases the intron, still in the form of a lariat.

IChanges in splice site choice arise from changes in the assembly of the spliceosome. The splice site consensus sequences are generally not sufficient information to determine whether a site will assemble a spliceosome and function in splicing. RNA elements that act positively to stimulate spliceosome assembly are called splicing enhancers. Exonic splicing enhancers are commonly found even in constitutive exons. Intronic enhancers also occur and appear to differ from exonic enhancers. Conversely, other RNA sequences act as splicing silencers or repressors to block spliceosome assembly and certain splicing choices. Again, these silencers have both exonic and intronic varieties. Some regulatory sequences create an RNA secondary structure that affects splice site recognition [][][], but most seem to be protein binding sites.

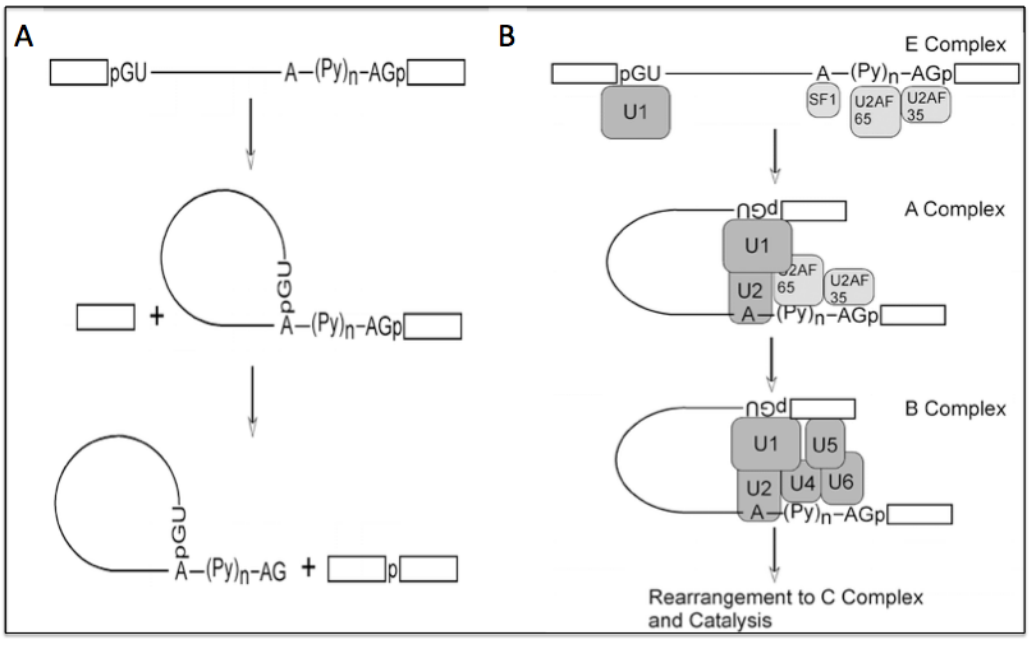


Figure 8 Splicing mechanism.

A. Splicing takes place in two transesterification steps. The first step results in two reaction intermediates: the detached 5′ exon and an intron/3′-exon fragment in a lariat structure. The second step ligates the two exons and releases the intron lariat. **B.** The spliceosome contains five small nuclear ribonucleoproteins that assemble onto the intron. The Early (E) complex contains the U1 snRNP bound to the 5′ splice site. Each element of the 3′ splice site is bound by a specific protein, the branch point by SF1 (BBP), the polypyrimidine tract by U2AF 65, and the AG dinucleotide by U2AF 35. This complex also apparently contains the U2 snRNP not yet bound to the branch point. The A complex forms when U2 engages the branch point via RNA/RNA base-pairing. This complex is joined by the U4/5/6 Tri-snRNP to form the B complex. The B complex is then extensively rearranged to form the catalytic C complex. During this rearrangement the interactions of the U1 and U4 snRNPs are lost and the U6 snRNP is brought into contact with the 5′ splice site. Adapted from Black [Black et al., 2003].

1.3.2 Alternative splicing and proteomic complexity

When an alternative splicing event occurs within the protein-coding region of a gene, it can modify the protein product in a variety of ways (Figure 9). The most common ones are “frame-preserving” and “frame-switching” alternative splicing events: in the first case the AS insert removes a peptide segment without affecting the rest of the encoded protein. In the second event, AS will shift the downstream reading frame of the protein. Genome-wide analyses of alternative splicing in human and other eukaryotes show that 40% of AS events are frame-preserving but interestingly, this percentage is much higher in evolutionarily conserved alternative exons. It is worth noting that a large number of frame-switching alternative splicing events introduce premature termination codons (PTCs) into the transcript isoforms []. The mRNA nonsense mediated decay (NMD) pathway is activated, leading to the degradation of the premature transcripts. NMD is a surveillance mechanism that detects and degrades mRNAs with premature stop codons. Importantly, more than a third of reliably inferred alternative splicing events in humans result in mRNA isoforms with premature stop codons (Hillman et al., 2004). The fact that this phenomenon is so widespread indicates that NMD does not necessarily have a function to prevent protein mistranslation when errors occur, but could also be a regulatory mechanism

that silences gene expression at post-transcriptional level.

A number of studies have focused on sequence analyses of full-length protein isoforms to investigate the global impact of alternative splicing on protein domains. Short alternative splicing within protein domains tend to target functional residues more frequently than expected by random chance. These data suggest that natural selection favors the use of AS in creating functional diversity of the proteome.

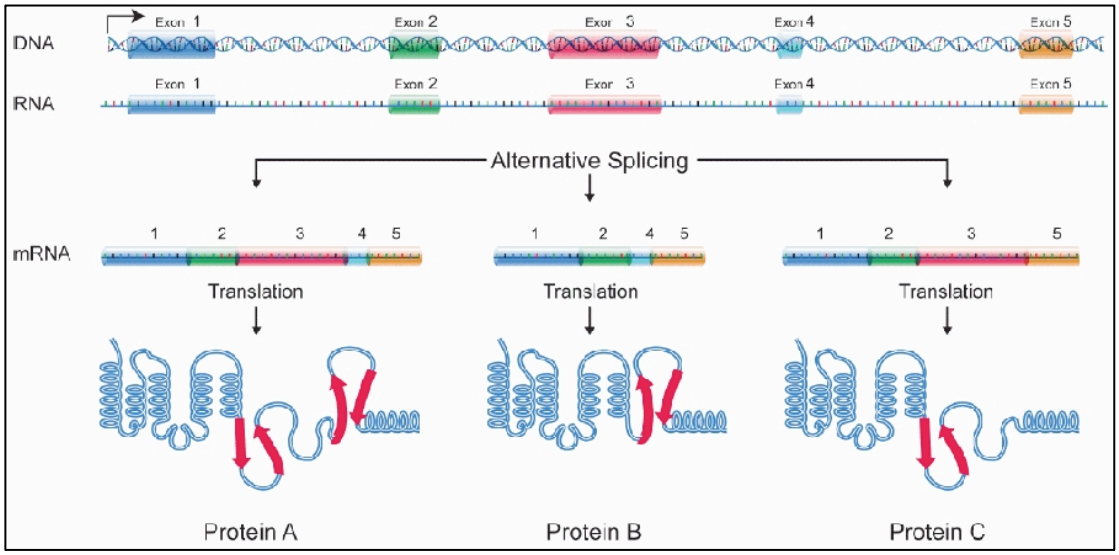


Figure 9 Alternative splicing increases the diversity of proteome.

Alternative inclusion of exons 3 and 4 in this example can change the structure and function of the resulting protein products. Adapted from Clancy [Clancy, 2008]

The impact of AS on protein domains represents an interesting issue in the regulation of transcription factor genes (TFs). Transcription factor genes are highly modular in the organization of sequence elements required for DNA binding, dimerization, ligand binding, subcellular localization and transcriptional activation, and a wide range of alternative splicing strategies operate on this modularity to generate different TF isoforms.

1.4 Gene duplication versus alternative splicing

Over the last few years, another major theme has emerged: the effect of alternative splicing on various processes of genome evolution, ranging from the small scale (individual nucleotide mutations) to the large scale (i.e. exon creation and loss).

Kopelman et al. (2005) found an inverse correlation between the extent of alternative

splicing and gene duplication in human and murine gene families. A similar anti correlation was observed also in *C.elegans* (Hughes and Friedman, 2008). The phenomena underlying the apparent mutual exclusion between the two competing mechanisms have thus far not been explained completely. The propensity toward expansion through duplication or alternative splicing is sometimes different for the same gene family in human and in mouse, so the inverse correlation seems not to be caused by inherent properties of the different gene families. This is remarkable since alternative splicing and duplication are very different processes with potentially different impacts on the functional differentiation of genes. As described in the previous paragraphs, alternative splicing tends to influence protein structure more drastically than duplicated gene divergence (Talavere at al., 2007; Shakhnovich and Shakhnovich, 2008). Also, differential regulation of splice variant expression may differ substantially in mechanism and effect from expression divergence of duplicated genes. One mechanism that has been invoked to explain the anti correlation between gene duplication and alternative splicing is the dosage balance effect: alternative spliced genes would, upon duplication, give rise to multiple additional isoforms that could exacerbate balance effects (Talavera et al., 2007). The evidence for the influence of dosage balance effects on duplicate gene (or different gene isoforms) retention is overwhelming. A simple interpretation of the cause of these transacting dosage effects is that the latter are caused by a gene or genes that exhibit a dosage effect themselves and that act in a regulatory fashion to modulate many targets. These genes are called dosage-sensitive genes and the most representative are transcription factors genes (TFs). As mentioned in the first paragraph, dosage-sensitive genes are significantly over-retained following WGDs and, in the same time, exhibit lower retention rates following smaller scale duplications (e.g., local and tandem duplicates, segmental duplicates, aneuploidy). In fact, a modification in the relative abundance of subunits in a transcription factor complex may alter the assembled complex and the expression of target genes (Birchler et al. 2001). Since transcription are among the most interesting examples of dosage-sensitive genes also because of their critical roles in gene regulation, I studied the organization of transcription factor gene families in a highly duplicated, reference genomes and I analyzed how this class of genes are regulated in terms of alternative splicing. In the next paragraph, I will provide an overview of transcription factor genes, in particular in terms of their duplication and their regulation through alternative splicing events. Chapter 4 describes how TF

families are classified and organized in plant, while Chapter 5 provides a description of human transcription factors and the dynamic usage of the different TF isoforms.

1.5 Eukaryotic transcription factors: key regulator of the transcriptional machinery

Transcription factor genes (TFs) are considered as key regulators of the transcriptional machinery. TFs are typically classified by their DNA-binding domain (DBD) type. Eukaryotic genomes display remarkable diversity in their transcription factor (TF) repertoires, in terms of both presence and prevalence of different TF families in different lineages. It is estimated that TFs constitute between 0.5% and 8% of the gene content of eukaryotic genomes, with both the absolute number and proportion of TFs in a genome roughly scaling with the complexity of the organism [1]. Most eukaryotic TFs tend to recognize short, degenerate DNA sequence motifs, in contrast to the larger motifs preferred by prokaryotic TFs [2]. Cooperation among TFs, rather than highly-specific sequence preferences, is believed to be a pervasive feature of eukaryotic transcriptional regulation [3]. The distinguishing feature of TFs, relative to other transcriptional regulatory proteins, is that they interact with DNA in a sequence-specific manner [4, 5]. In the vast majority of well-studied cases, these interactions are mediated by DNA binding domains (DBDs) [6], and TF families are typically defined on the basis of sequence similarity of their DBDs. Eukaryotic DBDs display a wide range of structural forms spanning a diverse array of protein folds (Figure 1), each of which represents a different solution to the problem of recognizing DNA sequences. Most strategies involve interactions with the major groove, although minor groove and/or phosphate and sugar backbone interactions also appear frequently.

1.5.1 Transcription factor DNA-binding domains

DNA-binding domains (DBDs) have been classified according to their three-dimensional structural properties. Basic description of the domains can be found in Latchman (2005). A more systematic and current classification of DNA-binding

domains was carried out by Stegmaier et al. (2004), in which DNA-binding domains were divided in superclasses, classes, families and subfamilies. According to this, five main structural superclasses can be distinguished:

-Basic domain
- Helix-turn-Helix domain
- Zinc coordinating domain
- β- scaffold with minor groove contacts domain
- other domains

Basic domains are characterized by a region rich in basic amino acid residues in alpha-helix conformation that can interact directly with the DNA. DNA-binding specificity is determined by the sequence of the basic region. This domain is usually accompanied by an additional domain, e.g., leucine zipper, helix-loop-helix or helix-span-helix, that does not interact directly with DNA, but that is important for dimerisation and for the correct positioning of the DNA-binding regions of the dimer. The helix-turn-helix domain consists of two alpha-helical regions arranged at right angles to each other. It has been shown that one of the two helices lies partly within the major groove of DNA (recognition helix), where the sequence specific interaction takes place.

In **zinc coordinating domains** the presence of zinc (Zn2+) is required for sequence specific DNA-binding. The zinc ion can be tetrahedrally liganded by either two cystein and two histidine residues (C2H2 zinc finger, not included in Stegmaier classification; STEGMAIER et al. 2004) or by multiple cysteine residues (C4 and C6 zinc fingers), allowing the formation of a structure called the zinc finger, which is responsible for sequence specific DNA-binding.

β- scaffold domains with minor groove contacts is a very diverse superclass, without a structural characteristic shared by all members. Their overall mode of interaction consists of inserting into the minor groove and causing a tight twist in the DNA.

In addition to the four major superfamilies a wide range of TFs employ less conventional strategies for recognizing DNA sequences, including the AT hook, which recognizes sequences in the minor groove utilizing fewer than a dozen amino acids [7], and the strongly twisted antiparallel β-sheet and four α-helices comprising the SAND domain [8]. These are grouped as other. Several excellent reviews have previously covered the topic of eukaryotic TFs [6, 9–11].

1.5.2 Non-DNA-binding Regions of Transcription Factors

Besides the DBD, TFs often present additional\accessory domains" involved in regulating the activity of the TF itself by mediating protein-protein interactions, activation potential or metabolite binding (Figure 2). As mentioned above, a TF's activity is often determined by context-dependent interaction with other regulatory components and indeed, many TFs are also capable of forming homo or heterodimers (Table 1.1; Luscombe et al., 2000). Especially in eukaryotes, heterodimerisation between TFs from the same family occurs often and provides mechanisms for combinatorial control. These include dimer- and hence context dependent DNA-binding specificity as well as additional regulatory mechanisms by providing concentration-dependent switches through stochiometrical requirements for certain complexes to form. Heterotypic interactions with binding partners that lack DNA-binding ability or activation potential can lead to the complete inactivation of an interacting TF through dimerisation with the latter (reviewed in Amoutzias et al., 2008) or concentration dependent effects where the inactive component acts as a molecular titer. Molecular titering in turn can result in ultrasensitive behaviour, where small changes in input concentrations, protein degradation rates or interaction strength of the partners involved can yield large changes in the concentration of the active TF (e.g. Buchler & Louis, 2008).

1.5.3 Gene duplication and transcription factors

Duplication of TFs is likely to be a major evolutionary force driving divergence in transcriptional regulatory networks. Once a TF is duplicated, one or both paralogs may be under relaxed selective constraint and accumulate mutations in the DBD and non-DBD regions, therefore becoming able to acquire new target genes or interactions with other regulatory proteins. This can allow subfunctionalisation, i.e. the partitioning of regulatory roles, or neofunctionalisation through the acquisition of new target genes under the same signal or the opposite scenario, altered control of the TF while maintaining its target genes thereby being able to integrate several signals. The genomic region that is duplicated during such events can vary in size from small-scale duplication of individual genes up to the duplication of entire genomes. Especially the

latter has been shown to be of great impact for regulatory evolution. After a WGD event on the yeast lineage, TFs were among the preferentially retained classes of genes and the post-WGD regulatory network has been shown to diverge rapidly in function trough asymmetric loss of regulatory interaction in paralog pairs of TFs (Byrne & Wolfe, 2007; Conant & Wolfe, 2006). WGD events can thus facilitate large, but most importantly, coordinated changes of entire regulatory programs through the creation of initial regulatory redundancy.

Duplicate copies of a TF will be redundant at first, removing strong selective constraints and allowing for sub- or neofunctionalisation of TFs through gain or loss of domains, interaction partners or changes in DNA-binding specificity. The fact that TF repertoires display such strong lineage-specific patterns of amplifications of DBD-types and domain architectures (Charoensawan et al., 2010b; see above) suggests that gene duplication is indeed a major evolutionary force driving the divergence of TF repertoires. Lastly, pleiotropic effects can be overcome through alternative splicing and use of different TF isoforms in different tissues.

1.5.4 Alternative splicing regulates TF's transcriptional activity

In addition to gene duplication, novel functionalities in TFs can be gained through other mechanisms such as alternative splicing. As described before, alternative splicing is one of the most important post-transcriptional mechanisms for the proteome diversity. Transcription factor genes are modular at the sequence level: one module corresponds to the DBD, while another is a regulatory domain that mediates gene activation. Alternative splicing acts on this modularity. Numerous examples have now been described in eukaryotes, where a single RNA from a particular transcription factor gene can be spliced in two or more different ways to yield different mRNAs encoding transcription factor proteins with different proteins (for review see Latchman 2005). For example, two alternatively spliced mRNAs are produced, one of which encodes the active form, while the other produces a protein lacking the DNA-binding domain (DBD). This truncated isoform is incapable to binding to DNA and activating gene expression [1] [2]. As well as affecting DNA binding specificity, alternative splicing can also produce forms of a transcription factor with different effects on transcription. For instance, alternative splicing can

produce two alternatively spliced mRNAs, one of which encodes the active form, whit either the DBD or the co-regulator domain, while the other produces a protein lacking of the co-regulator domain. This truncated isoform is capable to binding to DNA but unable to activate gene expression. The two isoforms compete for the same binding domain [3] [4].
In humans, ~95% of multi-exonic genes are alternatively spliced [11]: thus alternative splicing provides a potentially wide-spread and important mechanism for controlling the regulatory activities of TFs in different tissues [12]. So far, no study has examined the extent and impact of alternative splicing on transcriptional regulation in any mammalian genome. Without this information, we are unable to understand the gene-expression programmes that allow cells to take on their individual identities.

1.6 Outline of the thesis

The following chapters of the thesis present three distinct investigations. I first analyze the *Arabidopsis thaliana* genome in terms of duplicated and singleton genes (Chapter 2 and 3). In particular we implemented a bioinformatics pipeline to detect all the pair-wise paralogy relationships in the *Arabidopsis* genome. Moreover, the presented pipeline can provide a reference as tool for the detection of paralogy relationships in other genomes. Set of genes sharing one or more paralogy relationships were organized in networks, which are deeply described in chapter 2. The third chapter describes the analysis performed on *Arabidopsis* singleton genes, which presence in a so highly duplicated genome stirs up intriguing evolutionary issues. Both the analyses on duplicated and singleton genes lay a foundation for the work in the fourth chapter where *Arabidopsis* transcription factor gene families were analyzed. Integrating public data with the collections of networks and single copy genes, this work provides support to the classification of transcription factors in *A.thaliana* and represents a step forward to understand TF families organization and evolution. Transcription factors were also analyzed in terms of alternative splicing, using *Homo sapiens* genome as reference (Chapter 5). Analyses in Chapters 2 and 3 and Chapter 3 are being presented in at least two papers in preparation that will be submitted at the time when the thesis submission.

To further investigate on the impact of alternative splicing on transcriptional regulation, a genome-wide study of alternative splicing of TFs was also considered in the human genome, describing how alternative splicing affects TF isoforms in terms of regulatory domains. The last chapter is a brief analysis of the correlation between gene duplication and alternative splicing for both *A.thaliana* and *Homo sapiens*.

Thus, the overall aim of the thesis is the genome-wide investigation of two important biological mechanisms that provide the raw material for new biological functions. Since transcription factor genes represent an intriguing aspect in the field of gene duplications and alternative splicing as dosage sensitive genes, I used this class of genes as key example of the analyses here presented.

Chapter 2

Organizing the *Arabidopsis thaliana* genome in terms of duplicated genes

2.1 Introduction

In[i] 1996, when the plant science community decided to determine the genome sequence of the flowering plant *Arabidopsis thaliana*, few people suspected that this model plant organism were an ancient polyploid. Nevertheless, even before the completion of the genome sequence, it was clear that a large portion of its genome consisted of duplicated segments [Terryn et al., 1999]. After the analysis of bacterial artificial chromosome sequences, representing 80% of the genome, almost 60% was found to contain duplicated regions [Blanc et al., 2000], which strongly suggested a large-scale gene or even entire genome duplication events in the evolutionary history of Arabidopsis. This opinion was later shared by the Arabidopsis Genome Initiative, on the basis of the analysis of the complete genome sequence [*The Arabidopsis Genome Initiative,* 2000], and by Lynch and Conery [Lynch et al., 2000], who discovered that most Arabidopsis genes had duplicated approximately 65 million years ago (Mya), by using a dating method based on the rate of silent substitutions. Comparative studies between Arabidopsis and soybean [Grant et al., 2000] and between Arabidopsis and tomato [Ku et al., 2000] also suggested that one or more

large-scale gene or genome duplications had occurred. For example, in the latter study, two complete genome duplications were proposed, namely one dated 112 Mya and another 180 Mya, based on the presence of chromosomal segments that seemed to have been duplicated multiple times. The analysis of duplicated regions by the Arabidopsis Genome Initiative [*The Arabidopsis Genome Initiative,* 2000] did not reveal such segments. Vision et al. [Vision et al., 2000] also rejected the single-genome duplication hypothesis and postulated at least four rounds of large-scale duplications, ranging from 50 to 220 Mya. One of the age classes of duplicated blocks they defined ('100 Mya) grouped nearly 50% of all of the duplicated blocks, strongly suggesting a complete genome duplication at that time [Vision et al., 2000]. However, the dating methods applied in their study have been criticized [Wolfe, 2001]. A recent reanalysis of the duplicated blocks ascribed to different age classes, conducted by Raes et al. [Raes et al., 2002], indeed revealed that many of the ancient blocks described by Vision et al. [Vision et al., 2000] had a much more recent origin than was initially postulated.

It is clear that the discussion regarding the number and time of origin of large-scale duplications in *Arabidopsis* is far from being settled, partly because obtaining a complete picture of all duplications (and their dating) that have occurred in the evolution of a genome is not self-evident. Although the frequency of gene preservation over a large evolutionary period after duplication is unexpectedly high, and several models have been recently put forward to explain the retention of duplicates [Gibson T et al., 1998] [Lynch M., et al., 2000], [Wagner, A., 2002] [Gaut B. S., 2001] the most likely fate of a gene duplicate is nonfunctionalization and, consequently, gene loss [Lynch M., et al., 2000]. This observation has great consequences for the detection of duplicated regions in genomes. Identifying duplicated chromosomal regions is usually based on a within-genome comparison that aims at delineating collinear regions (regions of conserved gene content and order) in different parts of the genome. In general, one tries to identify duplicated blocks of homologous genes that are statistically significant, i.e., that are shown not to have been generated by chance. The statistics that determine collinearity usually depend on two factors, namely the number of pairs of genes that still can be identified as homologous (usually referred to as anchor points), and the distance over which these gene pairs are found, which usually depends on the number of ‘‘single’’ genes that interrupt collinearity [Gaut B. S., 2001] [Vandepoele, K., et al., 2002]. However, the

high level of gene loss, together with phenomena such as diploidization events, translocations and chromosomal rearrangements, often renders it very difficult to find statistically significant paralogous regions in the genome, in particular when the duplication events are ancient [Ku H. M., et al., 2000].

Two sequences are homologous if they share a common evolutionary ancestry. There are no degrees of homology; sequences are either homologous or not [Reeck GR, et al., 1997]. Homologous proteins almost always share a significantly related three-dimensional structure. When two proteins are homologous, their amino acid or nucleotide sequences usually share significant similarity. Thus, while homology is a qualitative inference (sequence are homologous or not), identity and similarity are quantities that describe the relatedness of sequences. Notably, two molecules may be homologous without sharing statistically significant amino acid (or nucleotide) identity. Recognizing this type of homology is almost a challenge in bioinformatics.

Proteins that are homologous may be classified as orthologous or paralogous. Orthologs are homologous sequences in different species that arose from a common ancestral gene during speciation. Paralogs are homologous sequences that arose by a mechanism such as gene duplication. Notably, orthologs and paralogs do not have the same function. In all genome sequencing projects, orthologs and paralogs are identified based on similarity searches. Two DNA (or protein) sequences are computationally defined as homologous based on achieving significant alignment scores. However, computationally defined homologous proteins do not necessarily share the same function, as it occurs in nature, were homologous sequences may acquire different functionalities. We can assess the relatedness of any two proteins by performing a pairwise alignment. Several alignment search tools exist that facilitate these analyses. However, in sequence comparison there is a tradeoff between sensitivity and selectivity. Two of the most popular tools used for sequence alignment are BLAST [Altschul SF., et al., 1990] and FASTA [Pearson W., 2004].

Based on the discussed issues, we considered the *A.thaliana* genome in terms of pairs of sequence related genes, with the aim to reliably define paralogs. We implemented a dedicated pipeline that takes into account several crucial aspects for the detections of duplicated genes. The duplicated genes were then arranged into networks where genes are included if they share significant similarity with at least another member of the network. Therefore network includes stuctural related genes that account for computationally defined paralogs. This is the reason why we defined the grouped

genes as network of paralogs, defining a specific organization of the genes of such a highly duplicated genome. The networks were made available to the scientific community for small and large scale dedicated analyses, through a web accessible database (available at http://biosrv.cab.unina.it/athparalogs/main/index). This resource may be useful either for evolutionary investigations or gene family analyses. Our works provides a novel viewpoint to the annotation of the Arabidopsis genome which may represent a valuable starting point for specific biological analyses, as the collection provided is useful for the investigation of gene families, the improvement of the annotation of unknown genes, the use of the Arabidopsis genome for comparative analysis for the study of other plant genomes, the understanding of evolutionary events.

2.2 Results

2.2.1 Duplicated genes: identification and networks organization

The main focus of our work is the re-organization of *A.thaliana* genome in terms of duplicated genes to improve its use as reference in plant genomics and to provide a tool for the investigation of its evolutionary history. To fulfill these aims, we performed a genome wide investigation of the *Arabidopsis thaliana* genome, identifying all the possible pairwise paralogy relationships among protein-coding genes and dividing them into two classes: duplicated and singleton genes. In this chapter the identification and the analyses of duplicated genes are described, whereas the results of the singleton genes analyses are explained in Chapter 3.

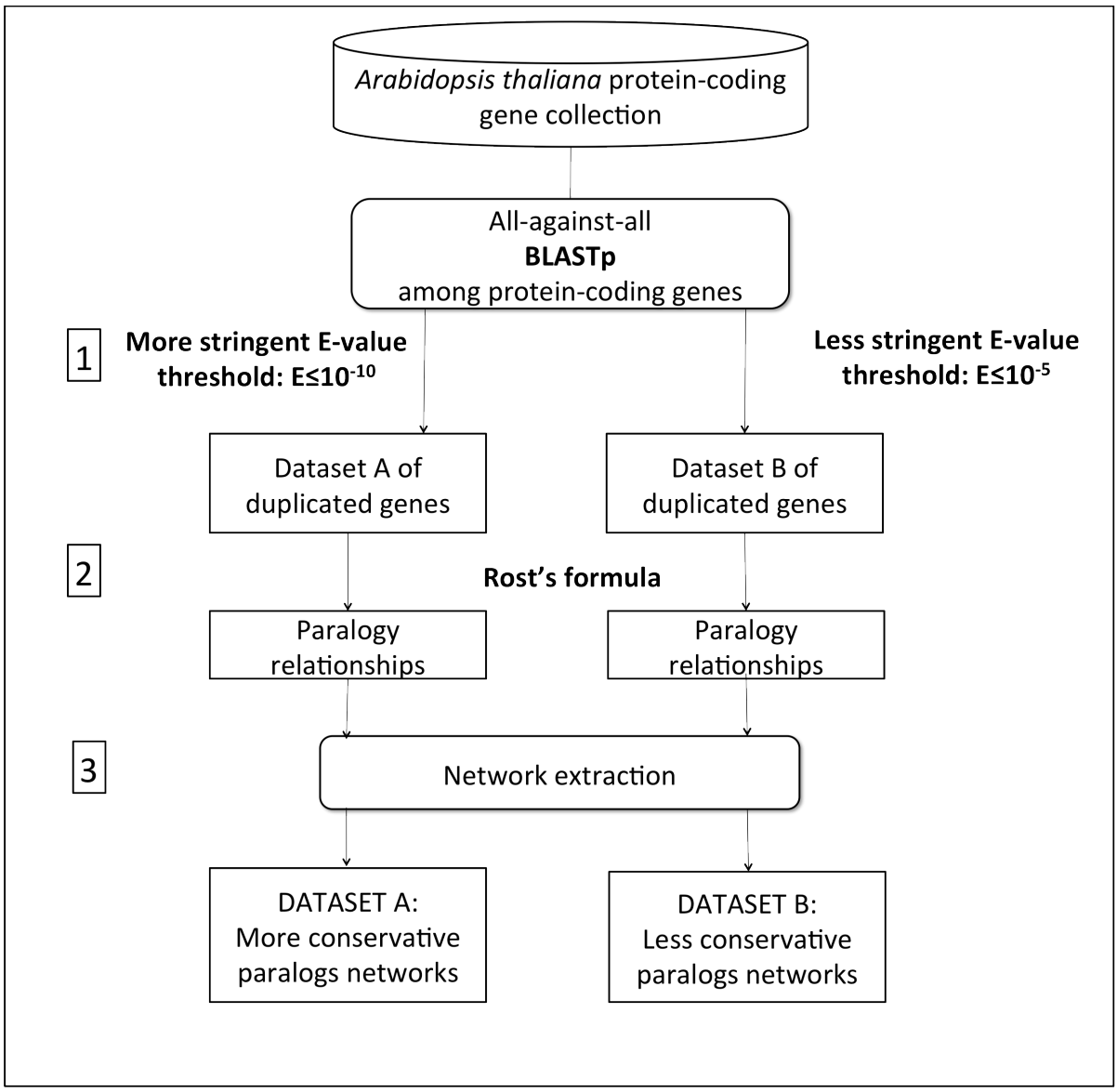


Figure 1. Pipeline for the identification of duplicated genes.

Starting from the *Arabidopsis* nuclear protein collection, all-against-all BLASTp analyses were performed to identify duplicated and singleton genes (step 1). Different parameters were considered producing two different datasets of duplicated genes (A, B). Duplicated genes were afterwards organized into networks (step 3).

2.2.1.1 Duplicated genes identification

In the first step of the pipeline we searched the 27169 *Arabidopsis* nuclear protein-coding gene collection from the TAIR 9 [The Arabidopsis Information Resource

(TAIR), 2009] for pairs of duplicated genes using an all-against-all BLASTp analysis of the encoded protein products. The analysis was repeated twice since two different E-value cutoffs were applied: a more stringent E-value threshold ($E \leq 10^{-10}$) and a less stringent one ($E \leq 10^{-5}$). Two different datasets of duplicated genes were found (dataset A and B, respectively) and in both cases about the 85% of the protein-coding genes show at least one significant sequence similarity versus another protein-coding gene (Table 1). We found that about the 85% of the associated genes show at least one significant sequence similarity versus another protein-coding gene (Table 1). Since defining the similar genes as paralogs is ambiguous if the alignments fall in the so called twilight-zone (20-30% of identities), we applied an empirical formula, the ROST's formula [Rost B., 1999] in the step 2 of the pipeline, in order to consider only reliable paralogy relationships. We identified those duplicated genes (405 and 113 genes, considering $E \leq 10^{-10}$ and $E \leq 10^{-5}$ cut-offs, respectively) with less than 150 amino acids aligned with an identity score lower than 30% (Supplementary materials). These genes were removed both from the networks and from the singleton collections due to the intrinsic ambiguity of their paralogy relationships (classified as "ambiguous due to the ROST's formula"). Duplicated genes obtained at this step (Table 1) were afterwards organized into the so called networks of paralogs (step 3 of the pipeline).

E-value threshold	Duplicated genes	Paralogs organized in networks
$E \leq 10^{-5}$	22927	22522
$E \leq 10^{-10}$	21956	21843

Table 1. Number of duplicated genes.

The number of duplicated genes are reported for each E-value threshold. In the third column the final lists of the duplicated genes included in the networks after removal of the genes filtered by the Rost's formula are shown.

2.2.1.2 Networks of paralogs

The networks can be described as groups of genes (the nodes of the network), with each gene connected to at least another one by one paralogy relationship, indicated as the edge of a network. It is worth to notice that, being each network a distinct connected component, every gene can belong to one and only one network. For each E-value threshold used, two distinct sets of networks were obtained: the more conservative ($E \leq 10^{-10}$) and the less conservative ($E \leq 10^{-5}$) ones. The first set consists of 3017 networks, while using the less stringent cut-off, 2754 networks were identified. The networks have different size, in terms of number of included genes, and complexity, in terms of number of paralogy relationships, as shown in the examples in Figure 2.

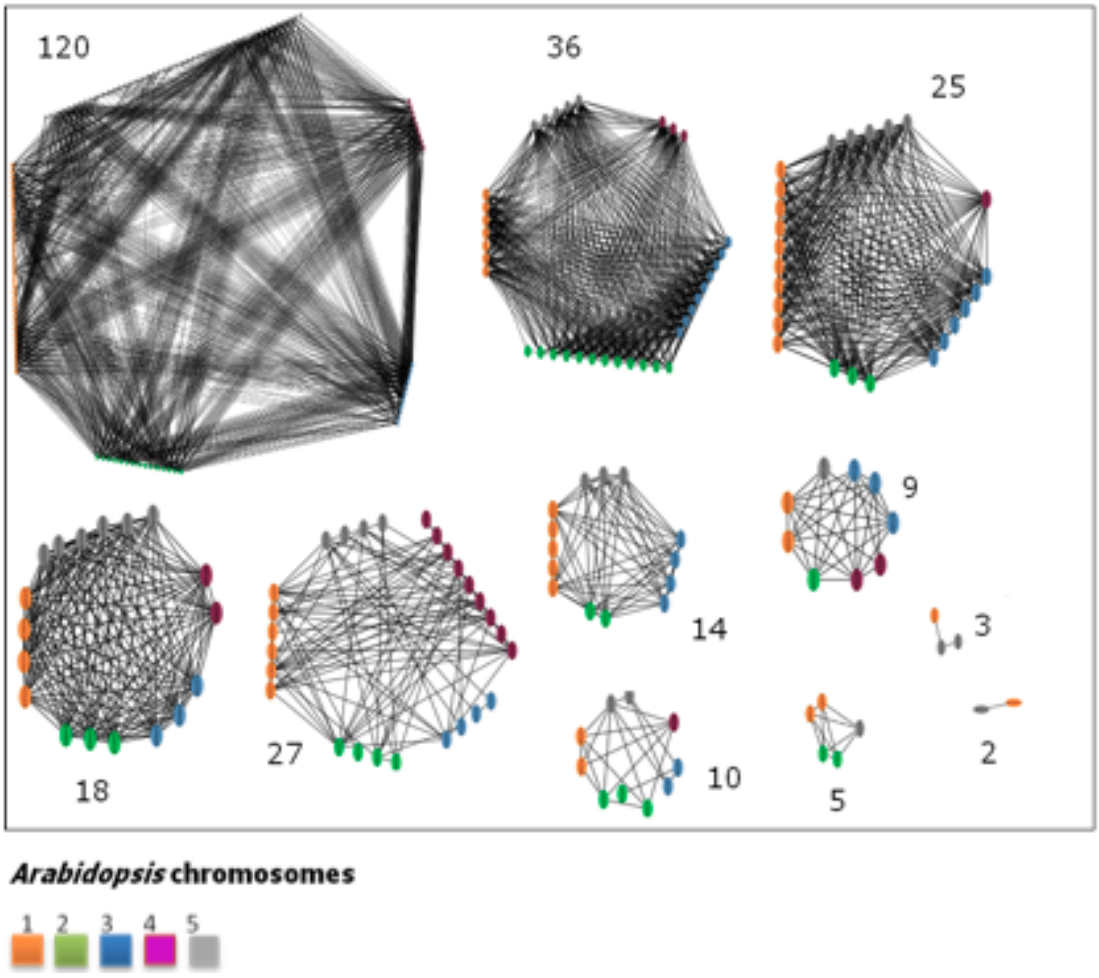


Figure 2. Network examples.

Differences in size and complexity exhibited from the extracted networks. Each oval is a gene, with the colors representing the five *Arabidopsis* chromosomes. Black lines represent paralogies between genes. Numbers indicate the number of genes per network.

In Table 2, the number of networks per size (i.e. in terms of number of genes involved) is reported for both sets.

Less conservative networks ($E \leq 10^{-5}$)		**More conservative networks ($E \leq 10^{-10}$)**	
No. of networks	No. of genes	No. of networks	No. of genes
1215	2	1347	2
1232	3 to 9	1370	3 to 9
241	10 to 30	216	10 to 30
65	31 to 209	83	31 to 207
1	210 to 6834	1	210 to 5168

Table 2. Network sizes.

For each E-value threshold, the size of the extracted networks with a given number of genes is reported.

About the 40% of the paralogs are distributed in small networks (2 to 9), the 17% of duplicated genes belong to medium size networks (10 to 30), while big networks (31 to 209) contain about the 18% of the paralogs. As a confirmation of the high complexity of the *A. thaliana* genome we found a huge network containing the 25% of the duplicated genes (more than 5000 nodes), spread over the five *Arabidopsis* chromosomes.

We further investigated on the number of intra- and inter-chromosome paralogy relationships for both the datasets ($E \leq 10^{-5}$ and $E \leq 10^{-10}$) (Table 3). Duplications are widely spread all over the five chromosomes, nevertheless it is possible to identify some preferential intra and inter chromosomal pattern of duplication. For instance, chromosomes 1 and 5 share the highest number of paralogies, as well as chromosome 1 is also connected with the chromosome 3 that is in turn extremely related with all

the other chromosomes.

A

		CHROMOSOME 1	CHROMOSOME 2	CHROMOSOME 3	CHROMOSOME 4	CHROMOSOME 5	No. PROTEIN CODING GENES per CHR
$E \leq 10^{-5}$	CHROMOSOME 1	35806					7054
	CHROMOSOME 2	33061	9697				4237
	CHROMOSOME 3	47078	25401	21208			5436
	CHROMOSOME 4	34927	18895	25670	12022		4124
	CHROMOSOME 5	52775	26755	39101	29768	23917	6318

B

		CHROMOSOME 1	CHROMOSOME 2	CHROMOSOME 3	CHROMOSOME 4	CHROMOSOME 5	No. PROTEIN CODING GENES per CHR
$E \leq 10^{-10}$	CHROMOSOME 1	33131					7054
	CHROMOSOME 2	30153	8695				4237
	CHROMOSOME 3	43110	23039	19335			5436
	CHROMOSOME 4	31941	17133	23474	11068		4124
	CHROMOSOME 5	48209	24151	35297	27141	21603	6318

Table 3. Duplicated genes distribution. The number of paralogy relationships among and within the different chromosome are here depicted. **A.** Results obtained using the less stringent E-value cutoff. **B.** Number of paralogies obtained using the more stringent .

2.2.3 Two-genes networks: an intriguing evolutionary issue

We detected a huge number of networks made by only two genes (hereafter called two-genes networks) representing the 10% of the genes in the genome (2694 genes with the more stringent E-value cut-off and 2430 genes with the less stringent one). Since the presence of genes with only one duplicate represents an intriguing issue in a so highly duplicated genome, we plotted the paralogies on the five chromosome to identify potential patterns of duplications (Figure 3) and we counted the number of pairs of duplicated genes among distinct regions of the chromosomes, indicating upper (U) and lower (L) arms of the chromosomes (Table 4). Also for this sample some favorite associations are detected. In particular, “stripes” of paralogies within the chromosomes (see for instance those within chromosome one in Figure 3), as well as inter-chromosome patterns (see those linking chromosomes two and four, as well as chromosomes one and three) are clearly depicted in the plot and summarized in the table 4.Moreover, the table shows that the highest number of two gene duplications are between the upper and the lower arms of chromosome 1, followed by the number of duplications between the two lower arms of chromosomes 2 and 3.

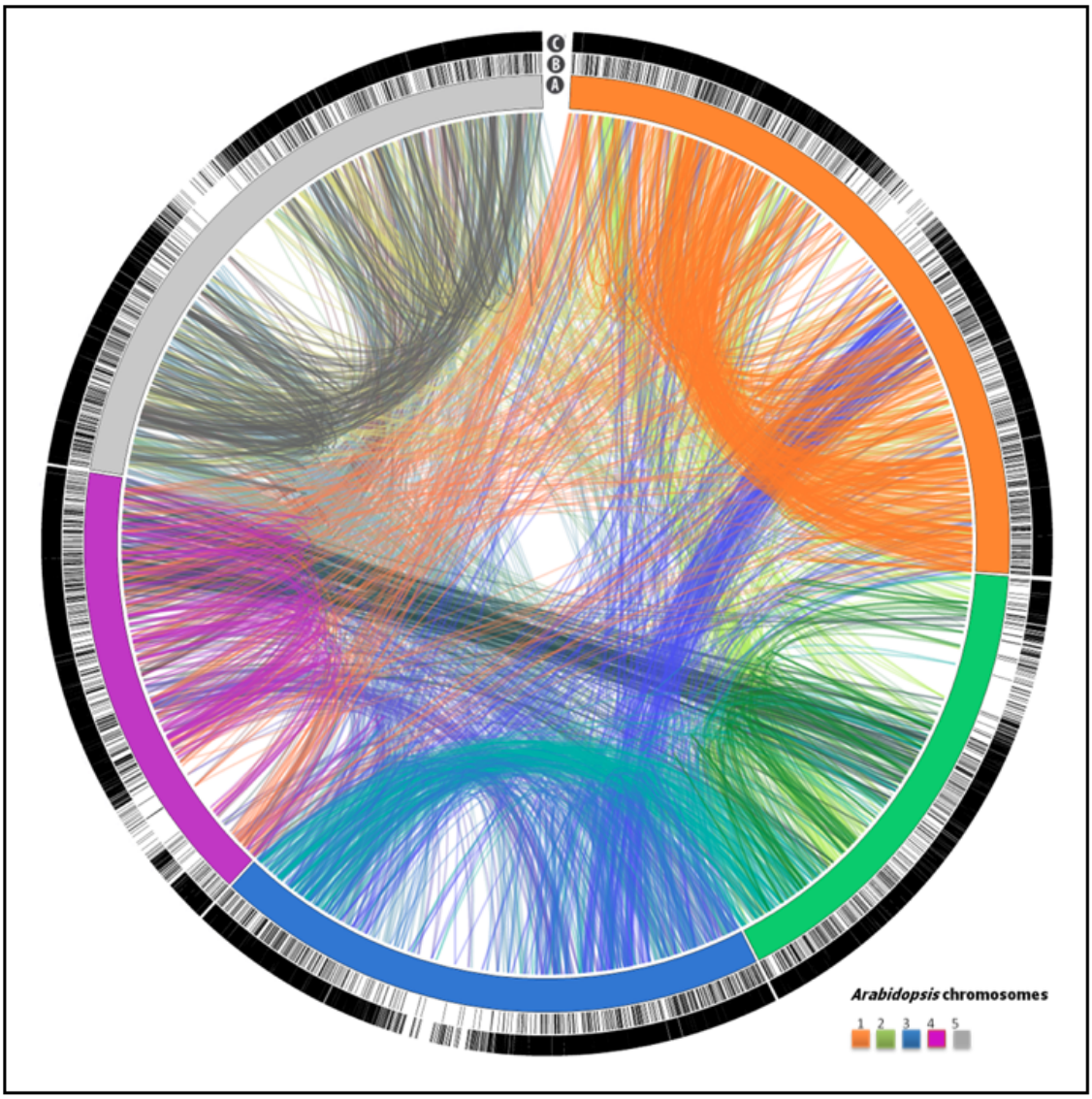


Figure 3. Two-genes networks The *Arabidopsis thaliana* chromosomes (solid colors). The lines included in the circle, linking the chromosome regions, indicate pairs of paralogs (i.e. two-genes networks) obtained with the $E \leq 10^{-5}$ threshold. Genes involved in networks of two genes are depicted in the circle B.

		CHROMOSOME 1		CHROMOSOME 2		CHROMOSOME 3		CHROMOSOME 4		CHROMOSOME 5	
		U	L	U	L	U	L	U	L	U	L
CHROMOSOME 1	U	21	94								
	L	94	29								
CHROMOSOME 2	U	7	4	6	1						
	L	61	19	1	22						
CHROMOSOME 3	U	20	53	4	20	22	13				
	L	17	15	0	80	13	7				
CHROMOSOME 4	U	14	4	0	2	5	5	2	5		
	L	31	24	2	64	34	13	5	34		
CHROMOSOME 5	U	16	17	2	21	71	21	6	16	29	36
	L	21	38	7	20	27	32	5	51	36	25

Table 3. Distribution of the genes in the two-genes networks. The number of genes shared between the upper (U) and lower (L) arms of the five chromosomes are indicated.

As a additional information, we calculated the nonsynonymous (*Ka*) and synonymous (*Ks*) substitutions for each pair of paralogy relationship within the two-genes networks to eventually detect some association of the intra- and inter-chromosomes duplications to their age (Figure 4A). No significant associations were found since different Ks and Kn values between genes are widely distributed among the strips, indicating that the inter and intra-chromosomes duplications of the *Arabidopsis* gene content occurred after the whole genome duplication events (Figure 4B).

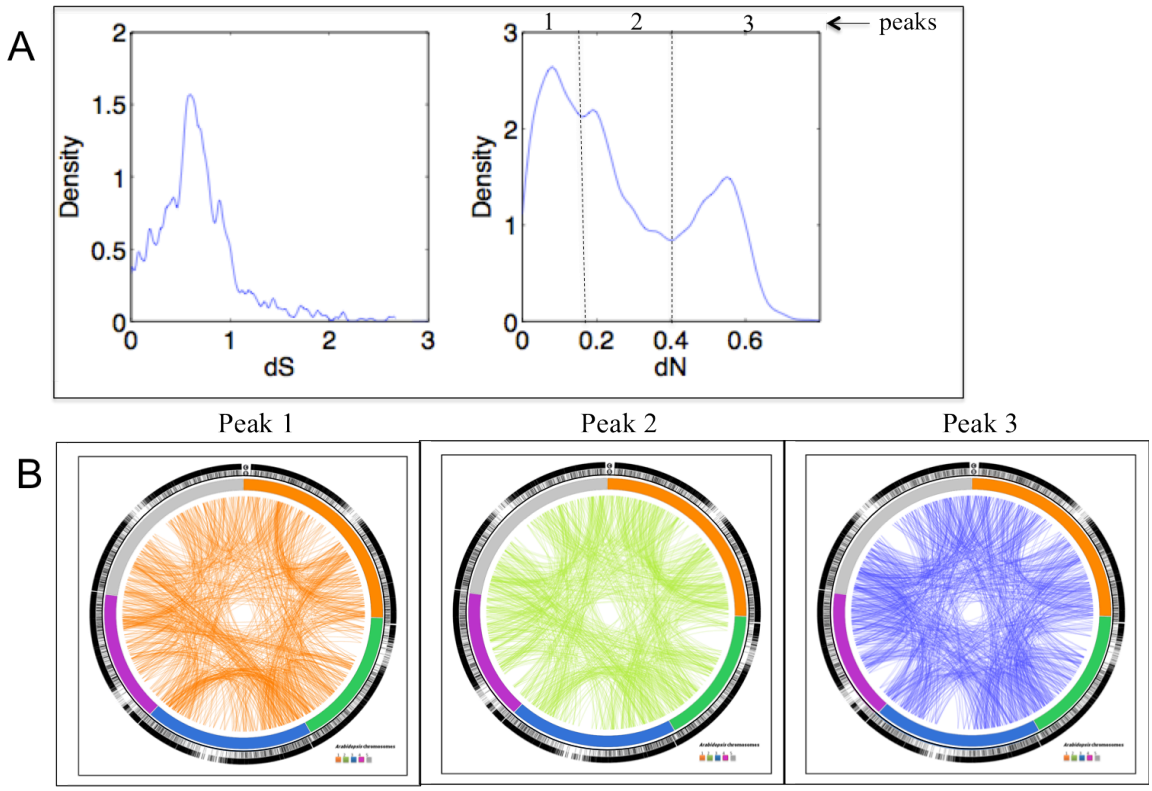


Figure 4. Two-genes networks evolutionary rates.

A. Plots of Synonymous (dS) and non synonymous (dN) substitution rates calculated for genes belonging to two genes-networks. We divided the dN plots in three peaks potentially correspondent to the three whole genome duplication events. **B.** For each circos image, two-genes networks with the same selective pressure acting on the involved protein-coding genes are depicted. In particular in orange, in green and in blue, paralogy relationships for which dN values belong to the first, the second and the third picks respectively are depicted.

2.2.4 More conservative and less conservative networks: some examples

The construction of networks of paralogs represents a useful starting point for disparate analyses: i) the investigation of gene families, ii) the improvement of the annotation of unknown genes, iii) comparative analysis for the study of other plant genomes, iv) the understanding of evolutionary events. The first aim is fulfilled mainly by the use of two different E-value cut-offs as shown in the example in Figure 5, in which genes belonging to two different families (a transcription factor family, C3H and the co-regulator gene family, BTB/POZ) are contained in the same network when the $E \leq 10^{-5}$ threshold is used, while they are split into two distinct ones, when the more conservative threshold is used ($E \leq 10^{-10}$).

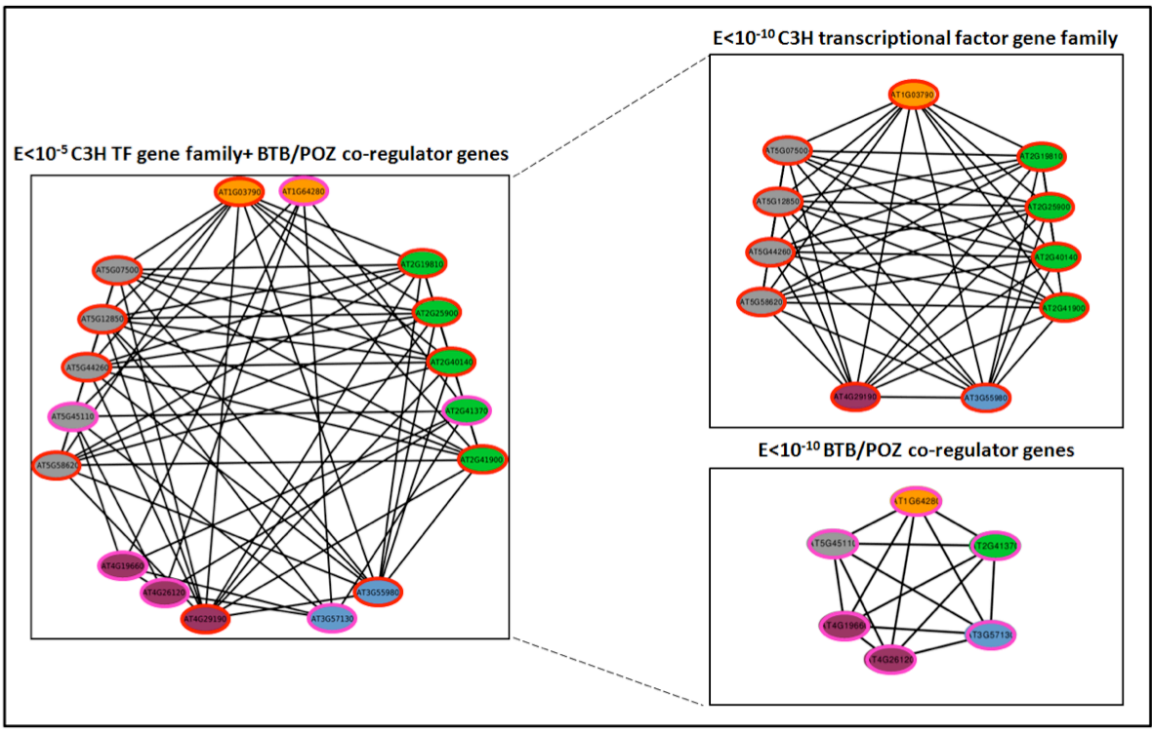


Figure 5. Network at different thresholds.

An example of different groupings due to the E-value thresholds. On the left one bigger network is obtained with the less stringent E-value cutoff. The same genes are split into two different networks when the more stringent threshold is considered. Red bordered genes belong to the C3H family, whereas pink circled ones belong to BTB/POZ family.

We compared the results obtained for the two different thresholds: 2501 networks remained unchanged in terms of contained genes, whereas 116 networks disappeared with the contained genes becoming all singletons when the more stringent threshold is used. The remaining networks were split into smaller distinct networks (as shown in Figure 5) or singleton genes. As an example, the largest network obtained at lower threshold $E \leq 10^{-10}$ is split into 116 smaller networks and 110 singletons when the threshold $E \leq 10^{-5}$ is used. See Table 2, in Supplementary materials, for details on how each of the $E \leq 10^{-5}$ network is modified when using the $E \leq 10^{-10}$ E-value cutoff.

The network in Figure 6B shows how structural relationships among genes with different functionalities can be highlighted: genes belonging to two distinct families (the transcription factor gene family MYB-related and the co-regulator genes belonging to the ARID family) are in t same networks together with genes not indicated to belong to any of the two families, thus suggesting relationships with different level of complexity that deserve further functional studies. The examination

of the networks can be a useful starting point for refining the unknown genes annotation, as shown in the small network in Figure 6A: it is composed by genes annotated as belonging to the same family, whereas the gene reported with a red circle is annotated as unknown. Since the unknown marked gene shares paralogy relationships with all the other genes within the network, this information can be the cue to further investigate and to finally assign the gene to the family.

Since *A.thaliana* is the reference plant genome, apart from the basic information related to gene structure and function (e.g., genome coordinates, mRNA and protein sequences, protein domains, and gene description), different types of genomics information are required to perform comprehensive comparative analyses. Our approach supplies a classification of the genome in terms of duplicated and singleton genes and represent a base to approach the study of new genomes as well as genome structure similarities within and between species.

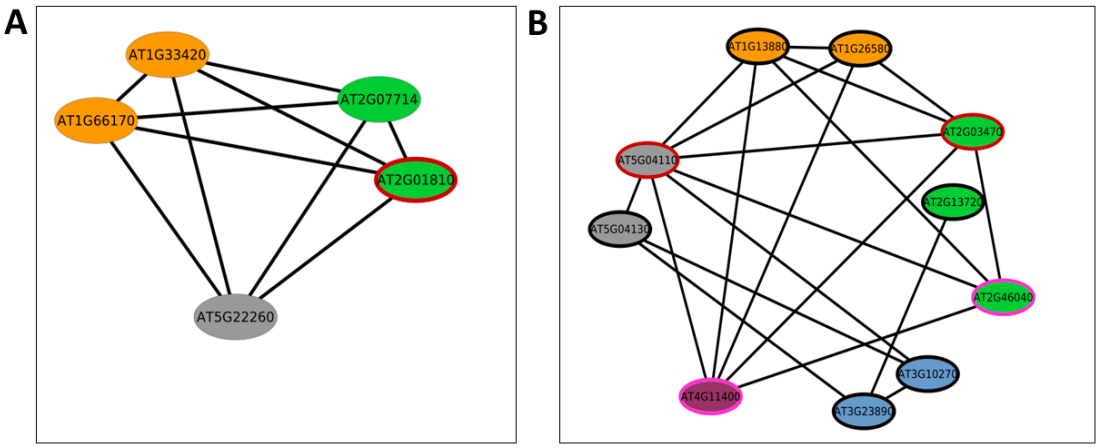


Figure 6. Network examples. A. The red bordered gene is annotated as an *unknown*. The others belongs to the same gene family. **B.** The red circled genes belong to the MYB-related family, the pink ones to the ARID family whereas the black ones aren't annotated as transcription factors.

2.2.5 TAIR9 versus TAIR10 releases

The recent release TAIR10 of the newest *Arabidopsis* annotation pushed us to verify our results on this collection. Since the number of new protein coding genes of the TAIR10 release is very small (lower than the 0.2% of the entire protein-coding gene collection), we based our validation on a qualitative comparison. We performed the

all-against-all BLASTp analysis applying the more stringent E-value cut-off ($E \leq 10^{-10}$) and then we organized paralogous genes into network. The number of genes having at least one paralog gene increased of less than one hundred units (representing the 0.35% of the total), due to an improved annotation quality. The size distribution of the networks remains substantially unchanged, as it is clearly shown in Table 4. A small shift from big sizes to smaller size of the network can be appreciated: the largest network is reduced of the ~3.7%, whereas the number of smaller networks increases with respect to the previous version. This trend is perfectly consistent with a revised version of the genome: indeed more reliable protein sequences imply narrower sets of paralog genes, hence corresponding to smaller networks.

A

Table 4. TAIR9 versus TAIR10

	More conservative networks $E \leq 10^{-10}$	
	Gene Number	**Network number**
TAIR9	2	1347
	3 to 9	1370
	10 to 30	216
	31 to 207	83
	208 to 5168	1

B

Table 4. TAIR9 versus TAIR10

	More conservative networks $E \leq 10^{-10}$	
	Gene Number	**Network number**
TAIR10	2	1346
	3 to 9	1232
	10 to 30	232
	31 to 209	68
	210 to 4948	1

Table 4. TAIR9 versus TAIR10.

Networks distribution in terms of size is shown for both the genome releases. **A.** The number of networks with a given number of genes obtained using the TAIR9 genome release is reported. The same information referred to the TAIR10 release is reported in **B.** Only more conservative networks results ($E \leq 10^{-10}$) are here shown. Similar results are obtained also comparing the less conservative networks ($E \leq 10^{-5}$).

2.2.6 Database construction and web interface: a genome resource

We used MySQL as the database management system and designed a uniform database structure for the obtained results. Either the more conservative or the less conservative networks extracted with our methodology as well as singleton genes' information (Chapter 3) are organized in the database that is accessible online at the address http://biosrv.cab.unina.it/athparalogs/main/index. The web resource we designed permits to obtain different information for each gene involved in the network analyses, including the list of direct paralogs (i.e. the genes directly related to the selected gene by a paralogy relationship), the list of all the genes included in the network, and the whole network itself as an XGMML file exportable for the Cytoscape environment [Smoot ME, 2011]. The use of different E-value cut-offs, permits to highlight several levels of similarity between same sets of genes.

2.2.6.1 Searching the database

In the query field it is possible to enter one of the following search keys:

Locus, whole or partial.

- RefSeq of encoded transcripts.

- Every string contained in TAIR Note field of locus annotation.

- Network name

It's possible to select via the drop-down list (near to the search field) one of the two e-value threshold used for the analyses. Hitting the Search button starts the query process. When the query process is ended, a list of one or more locus/loci is shown if matches are found. Otherwise a warning message is printed on the screen. On the top of the results table, the number of obtained loci is reported as well as the number of involved networks. Page numbers allow to browse all the results. It's possible to order (ascending/descending) the results by locus or number of paralogs or network name simply clicking on the heading. Each row of the list contains the following information:

- Locus: the gene locus id. Clicking on the locus is possible to browse the locus

details page (described in the next subsection)

- Paralogs number: the number of paralogs directly connected to this gene.
- Network name: the network containing the gene. Please note that each gene belongs to one and only one network. Clicking on the name permits to download the Cytoscape file (xgmml format) of the corresponding network.
- Number of genes in the network: the number of all genes contained in the network.
- Coding for: the transcript RNA type. Please note that only genes coding for mRNAs are included in the network analysis.

Encoded transcripts: the encoded transcripts RefSeqs. Each RefSeq is a link to the corresponding NCBI Gene detailed page.

Notes: the TAIR annotation for each of the RefSeq.

2.2.6.3 Locus detail and network visualization

Clicking on a locus name into the result table displays all its information in a new page. In particular: in the topmost part of the page, near the locus currently displayed, several lines summarize the information about the network and the locus details. The first line contains the chosen E-value and a clickable link that allows switching to the other threshold, updating the whole page. In the bottom left part of the page the list of locus paralogs is reported. Clicking on the link "Get list in plain text format" it is possible to download a text file with the list of all the loci. Clicking on a locus in the list updates the whole page with the selected locus details. In the bottom right part of the page a network excerpt is reported. Each oval is a locus, with the red-circled one representing the selected locus. Black lines are the paralogies connecting the involved genes. The color code is shown in the figure. Clicking on a locus within the network shows a new page including the selected locus details.

It's worth to note that for each gene of the network a pre-defined number of genes in the network is shown. This implies that, for big networks, only a small portion of the whole network is reported. To visualize the whole network it is necessary to download the xgmml file from the link situated on the top of the picture and import it

for a view within Cytoscape [Smoot ME, et al., 2010]. Clicking on the link "Download full locus list here" it is possible to download a text file with the list of all the loci included into the network.

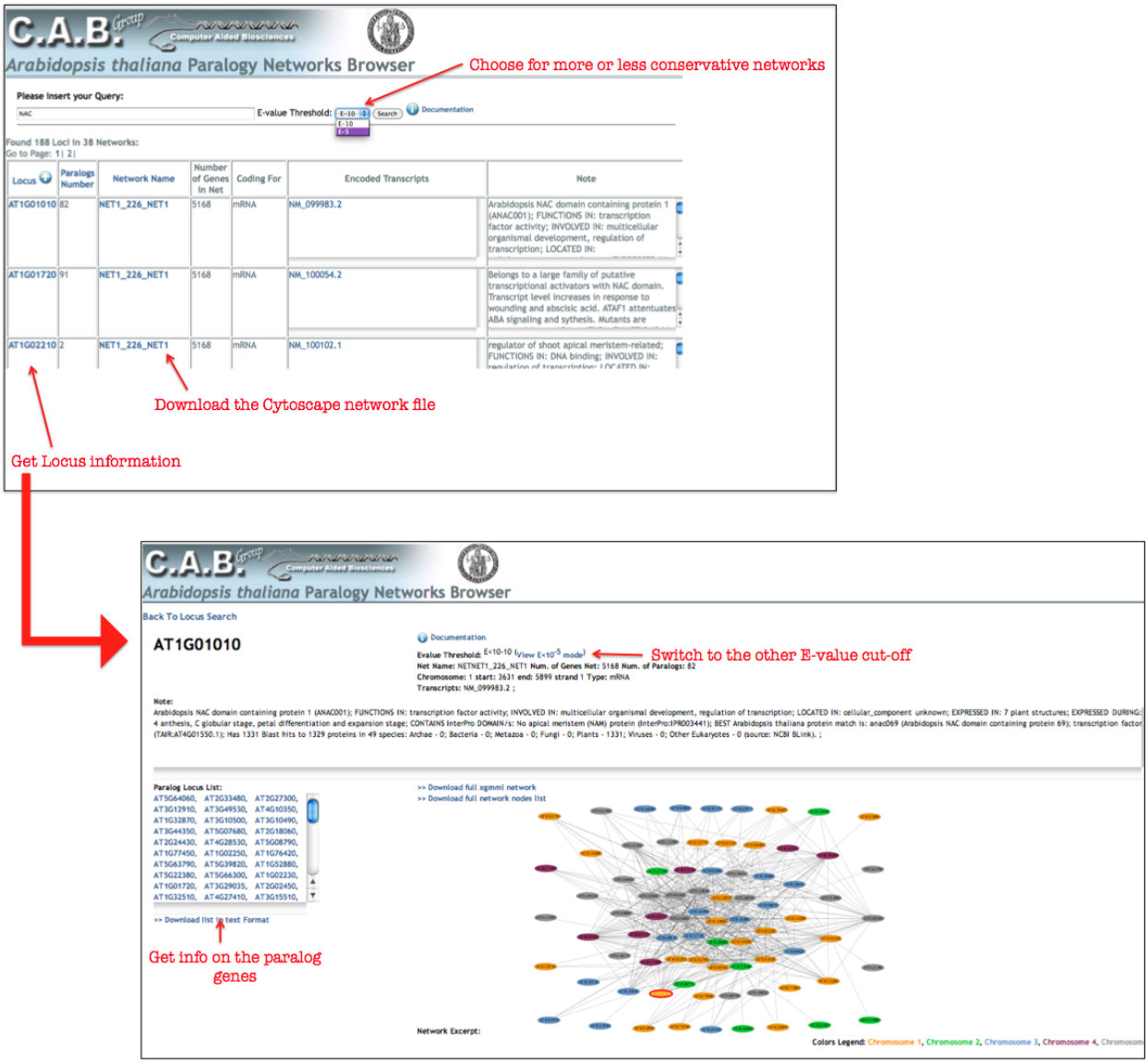


Figure 7. Different screenshots of the Arabidopsis thaliana paralogy networks browser.

Starting from the query page, the user can search the database for the name of the gene or for the gene family (in the example shown in this figure, the NAC family is used as query) using one of the two E-value thresholds ($E \leq 10^{-10}$ or $E \leq 10^{-5}$). The results of the query are summarized in a table shown in the panel above. Clicking on the name of the gene the user gets all the information about the chosen gene, the list of the genes in the network and the network in which the gene is contained. For big networks, only an extract of the network is shown, but it is possible download the entire network as a Cytoscape file (.xgmml) clicking

on the link above. It is possible to distinguish the query gene among the other genes in the network since it is red bordered. Clicking on the other genes (nodes of the network), the user can visualize the different connections within the network. In this page it is also possible to switch to the alternative E-value cut-off.

2.3 Conclusion and discussion.

The work described in this chapter highlights computational issues to be considered while producing sets of intragenome duplicated gene collections, evolutionary insights that come out from exploring such collections and the implications of our results in the assessment of the quality of the *Arabidopsis* annotation. Specifically, the main focus of our work is to highlight gene features of general interest when exploiting *A. thaliana* as a reference genome.

2.3.1 The importance of the right pipeline parameters

The design of the pipeline required for the proposed analysis faced well-known and still unexhaustively discussed issues related to the computational assessments of paralogs and to peculiarities of the specific methods employed [Van de Peer Y., 2004]. The issues also concerned the computational methods applied at each step and the parameter settings imposed. As an example, the alignment statistical significance, the presence of short sequences with identity percentage falling into the twilight zone and the presence of low complexity sequences have been properly considered to obtain reliable results and useful highlights on the structural relationships within and among the networks. For instance, though it is widely accepted that the most appropriate cutoff for defining paralogs within a genome is $E \leq 10^{-10}$, we considered that a less stringent threshold ($E \leq 10^{-5}$) could provide a coarse-grained view of the gene distribution into networks. In this frame, structural relationships among different gene families as well as among different class of genes, can be highlited by the use of two different statitiscal thresholds.

The Rost's formula [Rost B., 1999] allowed a better determination of genes which, though classified as “ambiguous due to the ROST’s formula”, cannot be ascribed as genes without copies in the genome.

The issue of implementing an adequate pipeline to classify duplicated genes first and obtain single copy gene then (Chapter 3) is not a novelty in genomics as well as in plant genomics [Duarte JM, 2010]. The methodologies are mainly based on blast supported approaches but, since the results are strongly affected by the specific steps followed and by the parameter setting applied, results may be difficult to compare, and it is compulsory to establish easily accessible results sources as reference for the community, avoiding the need of multiple similar efforts to collect data with similar requirements.

2.3.2 The advantages in using networks of paralogs

Ancient duplications and rearrangements of protein-coding segments have resulted in complex gene family organizations that may complicate the investigation on evolutionary and functional relationships within the families as well as of the genome organization and its evolutionary history. Extreme proliferation of some families within an organism, perhaps at the expense of other families, may correspond to functional innovations during evolution and our results may provide a results also within this frame. Over the last 10 years many papers tried to clarify the evolutionary history of A.thaliana [Vision et al., 2000] [Blanc et al., 2000] [Wolfe et al., 2001] but the number, the age and the importance of large duplication events remain still open questions. Studing Arabidopsis genome duplications in terms of “blocks” of paralogs [Vision et al., 2000][Van der Peer et al., 2005] or duplicated chromosomal segments [Blanc et al., 2003] based on protein sequence similarity is widely exploited as well as attempts to estabilish the timins of the events traced in the genome. It is not aim of this work to accomplish the contribute to the numerous efforts undertaken to uncover the events shaping the Arabidopsis genome. Our main interest was to set up of a reference framework and platform that could make the exploitation of networks of structurally related genes useful to establish intra and inter family and chromosome relationships. The organization of duplicated genes in networks helps is the elucidation of the organization of the Arabidopsis thaliana gene content in terms of pair-wise paralogy relationships. Moreover, their analysis highlights several evolutionary and functional issues that represent the basis for further investigations. In fact, neworks represent a first step to investigate genes and/or gene families structural relationships and evolution. Furthermore, our results allow to make a full

use of *Arabidopsis* as reference plant species, since the understanding of *Arabidopsis* genome organization is mandatory when it is used as model for other plant species analyses. Indeed gene organization into networks of paralogs is a valuable tool for genome analysis. First of all, networks can support the annotation process: in the TAIR 9 genome about the 19% of the protein-coding genes is still reported as unknown, and half of these genes belongs to networks. The annotation of either unknown genes or genes belonging to unknown family could be inferred from their paralogy relationships with other known genes presents in the network or belonging to specific family. Furthermore networks can help in highlighting the relationships, in terms of paralogy, of genes pertaining either to the same known family or to different related ones. The usefulness of networks of duplicated genes as a tool for genome investigation is illustrated in Figure 6. The newest release of the *A. thaliana* genome annotation required validation of the results obtain using TAIR 9. In particular, no big differences were shown. Specifically, it could be noticed that the number of two-genes network doesn't appreciably change using the newest release of the genome (Table 4).

The results can be accessed through an user-friendly and intuitive database. Here the distribution of genes among networks can be investigated in several ways. Is possible to search for a single gene locus or for a keyword contained into the annotation. For each locus is possible to know the number of its paralogs, the name of the network and the total number of genes herein contained. Other information such as the gene type, the transcripts codes (clickable and directly linked to the NCBI site) and the TAIR notes are alse available. By clicking on the network name is possible to download an Cytoscape file (xgmml format) of its description. By clicking on a locus name is possible to navigate to a detail page containing several information: first of all, a graphical view of the network (or of a portion of it when the number of contained genes exceed 150) is avalilable. Is possible to access the information on the other genes contained into the network simply clicking on the graph nodes representing the genes. All the descripted results can be obtained both for the E^{-10} and for the E^{-5} thresholds. Please note that all those genes excluded from the networks are marked as not included in network analyses.

This tool allows the scientific community to use our results for different purpouse and

researches, contributing to the understanding of plant genomics. In particular, this is the first web resource, in our opinion, that gaves the opportunity to view genes in the context of network of paralogous. This novel view of paralogy relationship between genes can be an helpful tool for researchers working in genome annotation and gene family studies.

2.3.3 Some evolutionary insight

Our analysis highlights the presence of different levels of complexity in terms of protein sequence similarity. In fact, we found a large network containing approximately one fifth ($E \leq 10^{-10}$) and one fourth ($E \leq 10^{-5}$) of the entire Arabidopsis protein complement, respectively. This evidence suggests a high degree of similarity of a large portion of the protein coding genes in this genome, maybe hiding a common evolutionary history, that surely will need further investigations to be assessed. On the other side, we found that about the 10% of the protein-coding genes has only a single copy in the genome, belonging to networks made of only two genes (two-genes networks), both for the less stringent and the more conservative thresholds. Considering the evolutionary past of A.thaliana, the high number of two-genes networks represents a rather intriguing aspect in a widely duplicated genome. Even more interesting is that genes involved in two-genes networks are equally distributed along the five chromosomes (Figure 3), suggesting consistent chromosome reshuffling after the WGD events or segmental duplication events.

2.4 Material and methods

2.4.1 Data retrieval

The *Arabidopsis thaliana* genomes (release TAIR9, June 2009 and TAIR 10, November 2010), were downloaded from the TAIR website (http://www.*Arabidopsis*.org/). TAIR9 consists of 27,379 protein-coding genes, 4827 pseudogenes or transposable elements and 1312 ncRNAs. TAIR10 consists of 27,416 protein coding genes, 4827 pseudogenes or transposable element genes and 1359 ncRNAs. We removed from protein coding genes the mitochondrial and chloroplastic ones, reducing the TAIR9 and the TAIR10 datasets to 27,169 and to 27206 genes,

respectively.

2.4.2 The pipeline

In the first step of the pipeline, an all-against-all protein sequence similarity search was performed using the BLASTp program [Altschul S.F., 1997]. BLAST or Basic Local Alignment Search Tool, compares protein and DNA sequences. BLAST begins its search process by comparing the query sequence with the entire database in search of *maximal segment pairs*. The maximal segment pair is the highest scoring pair of identical length segments chosen from two sequences. This maximal segment pair is a measure of local identity between two sequences. This segment pair is locally maximal if its score cannot be improved by extending both segments. BLAST reduces computation time by limiting results to those consisting of segment pairs whose score is above some threshold *T*.

An exhaustive analysis between all *Arabidopsis* protein sequences (27,369) was performed. Since the BLAST comparison of sequence A with sequence B is roughly equivalent to the comparison of sequence B with sequence A for significant hits, thus, just less than one half of the comparisons are redundant and only comparisons should be made. Moreover, we ignored the identity comparison of sequence A with itself, thereby minimizing an exhaustive N^2 comparison to comparisons.
To obtain a reliable determination of protein paralogy relationships we considered different criteria, described in the next substections.

2.4.2.1 The E-value cut-off

Among the absolute criteria that can be used to decide whether genes are true homologues (or in this case paralogues), the Expected value (E-value) is one of the most important one. It is the statistical significance threshold for reporting matches against database sequences; the default value is 10, such that 10 matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance ascribed to a match is greater than the expected threshold, the match will not be reported. Lower E-value thresholds are more stringent, leading to fewer chance matches being reported. The BLASTp analysis was based on two different cut-offs: a more stringent threshold, where the E-

value is $E \leq 10^{-10}$, and a less stringent one, with $E \leq 10^{-5}$ [He X, 2002] [Rubin GM, et al., 2000].

2.4.2.1 The twilight zone

Predicting whether two proteins are paralogous is relatively simple when their sequence identity (*I*) is high (>40% for long sequences) but becomes difficult when *I* is in the medium range (20–35%) or lower, especially for short sequences. Pairwise sequence identity (percentage of residues identical between two proteins) is not sufficient to define the twilight zone. Rost [Rost B, 1999] proposed an empirical formula for clustering proteins in a database (Table 4). Two proteins are assumed to be paralogous if the proportion (*p*) of identical residues over the *L* aligned amino-acid residues between the two proteins is higher than the cut-off point (pI) defined by the formula. The cut-off point increases as *L* decreases because two unrelated short sequences may by chance have a high *p* value. A common practice in clustering proteins into groups is to use single linkage: if proteins A and B have a *p* higher than pI and so do proteins B and C, then A, B and C are clustered in the same group, even if the *p* value for A and C does not meet the cut.

We applied Rost's formula with $n = 5$ (*n* is a factor to raise the cut-off point) to the aligned protein sequences. This cut-off was chosen according to Li et al.: they propose to use $I' = I \times \text{Min}\ (n_1/L_1,\ n_2/L_2)$, where *I* is the proportion of identical amino acids in the aligned region (including gaps) between the query (sequence 1) and target (sequence 2) sequences obtained by the alignment, L_i is the length of sequence *i*, and ni is the number of amino acids in the aligned region in sequence *i*. The factor Min $(n_1/L_1,\ n_2/L_2)$, which means the smaller of n_1/L_1 and n_2/L_2, takes care of the situation where a high *I* value is obtained when a short protein shares one or more domains with a longer protein. Another difference between I' and p' is that I' imposes a gap penalty in the aligned region. For short proteins, however, I' may become high by chance and so we impose $I' \geq pI$ with $n = 5$.

2.4.3 Networks extraction and visualization.

The network extraction process is based on considering the whole *A. thaliana* protein-coding gene collection as an undirected graph with vertex representing genes and

edges representing paralogies between genes, when present. Networks correspond to the connected components of the graph and are extracted with a recursive depth first search on it. Networks can be visualized either by Cytoscape [Smoot et al., 2011] or Circos [Krzywinski et al., 2009].

Cytoscape is a standalone Java application. It is an open source project under LGPL license. Cytoscape comes with various data parsers or filters that make it compatible with other tools. Among the different file formats that are supported to save or load the graphs (SIF, GML, XGMML, BioPAX, PSI-MI, SBML, OBO) we used the xgmml one, downloadable through the website.
Circos is a visualization tool in which two or more genomes can be representes as arcs in a single circle. Tracks of a variety of types can be aligned as inner circles along the genomes. Lines cross the middle of the circle connecting aligned regions. In this case the lines within the circle depict paralogy relationships.

2.4.4 dN and dS estimation.

In order to estimate the synonimous (dS) and non-synonimous (dN) substitution rate for the paralogy relationships belonging to the two-genes networks, we used Matlab's *dnds* function. This function estimates the synonymous and nonsynonymous substitution rates per site between the two homologous nucleotide sequences, by comparing codons using the Nei-Gojobori method [Nei et al., 1986]. This analysis assumes that the nucleotide sequences are codon-aligned, that is, do not have frame shifts. Moreover, it excludes codons that include ambiguous nucleotide characters or gaps and it considers the number of codons in the shorter of the two nucleotide sequences. We used as input the fasta sequences of the coding sequences (CDS-nt) of the pairs of paralogy relationships within the two-genes networks obtained with the less stringent E-value cut-off ($E \leq 10^{-5}$).

Chapter 3

Single copy genes: an evolutionary intriguing issue in a highly duplicated genome

3.1 Introduction

In the previous chapters we discussed about the importance of genome duplication in the evolution of biological diversity. Nevertheless, all characterized genomes include single-copy (singleton) genes, i.e. genes without apparent homologs within the same genome [14] and, for some of them, without homologs, even in phylogenetically close relatives [15]. Indeed, evolution is not a one-direction process and a high proportion of duplicated genes are rapidly lost [6,16,17]. This definition of singleton genes is fully independent of the gene function and is only based on the sequence uniqueness in the whole gene content of a considered species.

Unlike gene duplication, gene loss is not an unspecific mechanism but it is instead influenced by selection [12,18]. Thus, duplicates that are maintained show a bias toward certain gene functional classes [19] or transcriptional levels [6,20,21].

Duplicated genes resulting from whole-genome duplication have longer life expectancies than duplicated genes arisen from smaller scale duplication [] and tend to reduplicate in new genome duplication events. By contrast, genes that are rendered singleton following a genome duplication tend to be repeatedly returned to the singleton status during successive genome duplications [] []. Also evidence from

wheat indicates that duplicated low-copy number genomic regions, which may include low-copy number genes similar to which we consider as singletons, are rapidly eliminated following polyploidization [47]. In the rare instances in which duplicated copies of single or low copy genes are maintained over long evolutionary periods (tens of millions of years), paralogs show distinct patterns of functional and/or expression divergence. For example, over expression of LEAFY, a plant specific transcription factor gene, generally results in early flowering [48,49] and cases in which LEAFY is present in duplicate (typically in recent polyploids), expression patterns are typically complementary, suggesting that subfunctionalization may be necessary for the maintenance of both loci [50,51]. Moreover, singleton genes may also be duplicates that diverged too much to be still distinguishable today [22].

With the recent availability of whole plant genomes it is possible to further consider some questions about the generation and the evolution of singleton genes. The discovery in angiosperms of functional genes that are "duplication resistant", i.e. which are preferentially returned to the singleton status following genome duplications, adds a new dimension to classical views that focus on the potential advantages of genome duplications as a source of genes with novel functionalities.

In this frame, there has been a significant amount of attention paid towards the prospect of identifying single copy nuclear genes in flowering plants, primarily for their potential use as phylogenetic markers [4-6,27-32]. A number of low copy number nuclear genes have been previously identified in flowering plants, including the phytochromes, ADH, TPI, GAP3DH, LEAFY, ACCase, PGK, petD, GBSSI, GPAT, ncpGS, GIGANTEA, GPA1, AGB1, PPR and RBP2, primarily useful as phylogenetic markers [5,32-46].

Gene loss complicates genome comparisons by fragmenting ancestral gene orders across multiple chromosomes and, together with rearrangements of duplicated genes, makes evolutionary and comparative studies really challenging.

In this chapter we define as single-copy or singleton gene, a gene coding for a protein without detectable sequence similarities with any protein in the same proteome first, and then also with the rest of the genome (i.e. intergenic regions, non protein-coding genes).

We first established and used stringent criteria in order to identify suitable sets of unique genes present in the extensively known protein collection of Arabidopsis thaliana. In a second step of the study we identify a more reliable list of singletons

removing those genes sharing sequence similarities with non-protein coding genes and intergenic regions, since we consider these genes as ones which had a copy but it's not functional anymore or it's no longer annotated as a gene. Another interesting aspect that arises in this chapter is the identification of some flows in the Arabidopsis genome annotation, suggesting the need of a more accurate annotation process for the Arabidopsis genome. The results presented, together with the two-genes networks identified and described in the Chapter 2, stir up intriguing evolutionary issues concerning the evolution of the Arabidopsis genes, as well as those related to the presence of genes with at most one paralogy relationship in a highly duplicated genome.

3.2 Results

3.2.1 Singleton genes identification

We identified single copy genes, i.e. genes without copies among the protein-encoded complement, performing the first three steps of the pipeline shown in Figure 1.

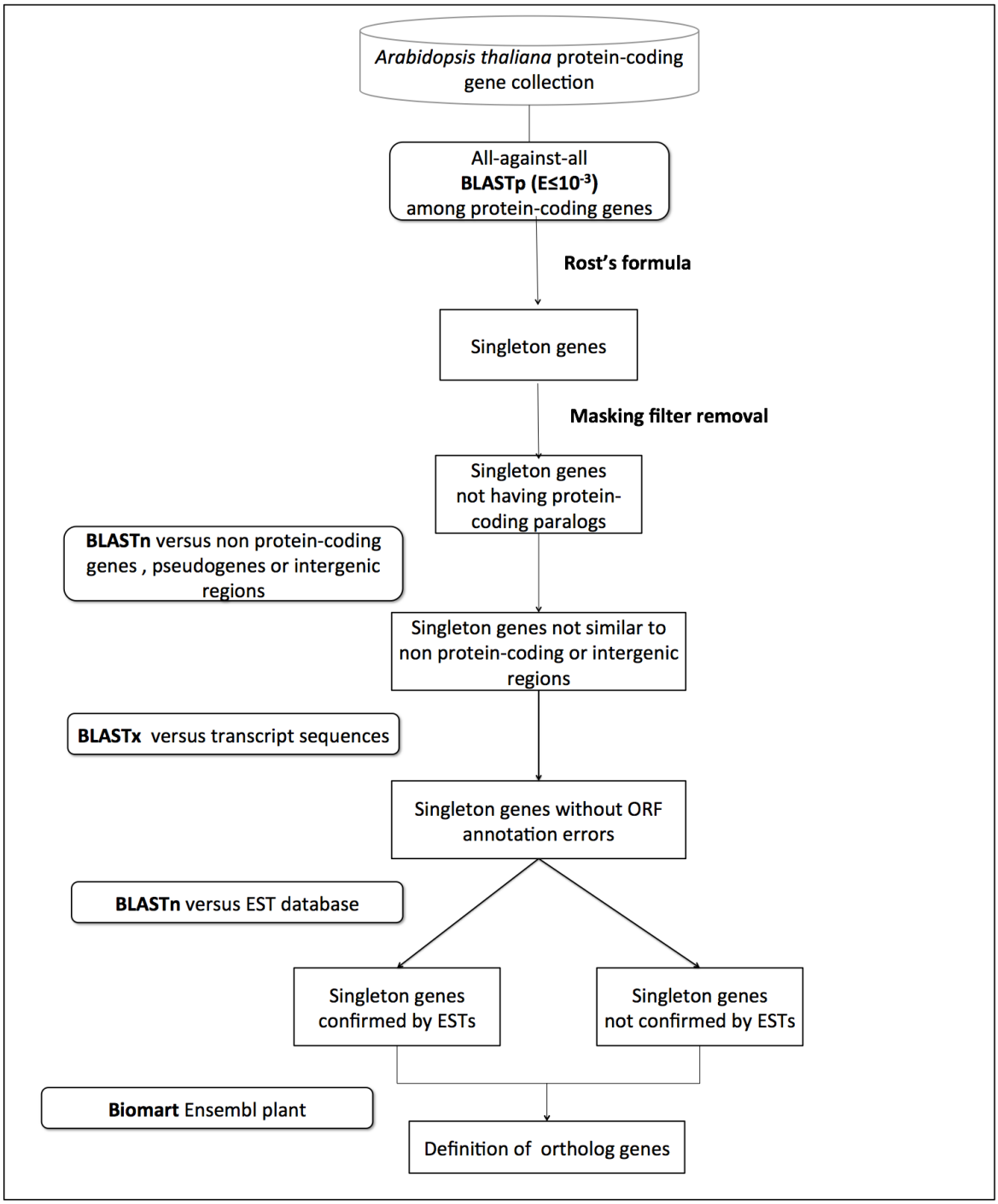


Figure 1. Pipeline for singleton genes identification and analyses.

The first two steps of the pipeline consist of an all-against-all Blastp analyses using a flexible E-value cut-off E≤10^{-3}. The obtained alignments are further filteres by the use of the Rost's formula as described in Chapter 2. The removal of the masking filter allowed the identification of true singleton genes, i.e. genes without copy within the protein-coding genes. The subsequent steps of the pipeline are aimed to better investigate on the single-copy genes: a Blastn analysis against the rest of the genome, allowed us to identify genes without any significant sequence similarity within the nuclear genome. These genes were there divided in

confirmed and not confirmed by ESTs. A comparative analysis with other plant species allowed the identification of singletons with at least one ortholog with the considered plant species (*O.sativa*, *V.vinifera*, *P.trichocarpa* and *S.bicolor*).

Basically, the first two steps are the same ones described in Chapter 2 for the identification of duplicated genes, but this time the analysis is focused on those genes not sharing sequence similarities with other sequences. The all-against-all BLASTp was performed, but in this case using a more loose E-value cut-off $E \leq 10^{-3}$. The use of a not stringent expected value threshold is aimed to identify traces of similarities between genes diverging more than those selected with a more stringent analysis. As for the identification of duplicated genes, the obtained alignments were further filtered by the Rost's formula, to consider those alignments with the 25-30% of identity (twilight zone). Again, in this case, the discarded genes (Suppementary table 1) were not considered for further investigations as singleton genes since the ambiguity of their nature. This approach allows the identification of 3802 genes not sharing significant sequence similarity with other proteins.

The analysis was also extended considering the default masking parameter of the BLAST tool (see Materials and Methods). In fact, one of the major problems in sequence homology searches is the presence of low-complexity sequences, also known as "simple sequence repeats", which have an unusual composition and are very abundant among the proteins [Golding, G.B. 1999]. In particular, due to the repetitive nature of these sequences, high scoring hits without biological meaning may be reported only because of the presence of a low-complexity region [Sharon I, 2005]. BLAST applies the SEG program [Wootton and Federhen, 1996] by default to filter out low-complexity regions in query protein sequences. But often true sequence similarity can be missed using the filter, in particular in the case of smaller sequences. So we first performed the BLASTp analyses with the SEG filter activated by default. Afterward, to recover genes that could be masked, therefore hiding similarities with other genes, the BLASTp analysis ($E<10^{-3}$) without the SEG filter's activation was performed. thus permitting to exclude 214 genes from the singleton list (Supplementary materials), determining a total amount of 3588 single copy genes. An example of paralogies missed by the masking filter is represented by the duplications within the DNA-binding S1FA protein family. According to the TAIR9 annotation, the family is composed of three genes: AT3G53370, AT3G09735 and AT2G37120.

The latter resulted as singletons when performing the BLASTp analysis with the E-value cut-off $E \leq 10^{-3}$ and the default masking filter. When the analysis was repeated and the masking was removed, the masking filter removal allowed the identification of genes sharing similarities.

Since the identification of thousands of single copy nuclear genes in the *Arabidopsis* genome raises interesting biological questions, we further investigated these genes as follows: i) identifying protein sequence similarities with the rest of the genome (non protein-coding genes and intergenic regions); ii) searching for possible annotation mistakes (ORFs); iii) confirming their protein-coding function using ESTs; iv) searching for orthologs through a comparative analysis. These steps are summarized in the pipeline in Figure 1, while the results of each analysis are hereafter described and summarized in Figure X at the end of the Result section.

3.2.1.1 Singleton genes analyses: searching for sequence similarity with the rest of the genome.

The sequence similarity search between the 3588 singleton genes and the remaining part of the nuclear genome, i.e. non protein-coding genes and intergenic regions, using a BLASTn analysis ($E \leq 10^{-5}$) resulted in a list of 3410 genes, hereafter considered as "true singleton" genes, as they share no similarity within the entire genome. Specifically, 250 shared a significant similarity with at least one *Arabidopsis* non protein-coding gene (Supplementary materials). These alignments were further investigated, taking into account the intron-exon structure of the singletons: only alignments involving exonic regions and where the match length was greater than or equal to 50% of the single copy gene transcript length, were considered as positive similarities. According to these criteria, we classified 178 genes out of 250 as genes with one or more non protein-coding paralogs, whereas the remaining 72 were confirmed as singleton genes (Figure 2). The Venn diagram in figure 4 B shows the distribution of the 178 genes among different classes of non protein-coding genes considered (see Material and methods).

No significant hits were found when singletons were aligned against intergenic regions, excluding the possibility that those genes had paralogs not annotated as gene.

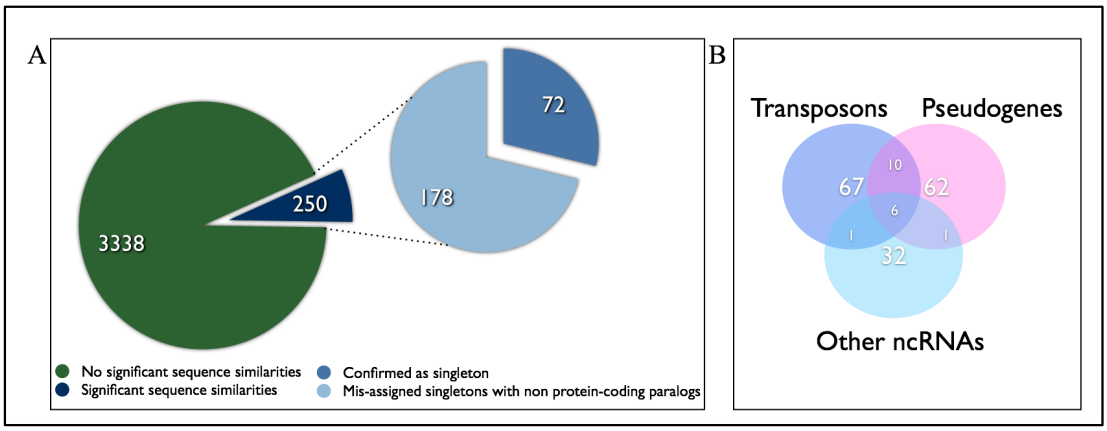


Figure 2. Sequence similarity search with non protein coding genes.

A. Singleton versus non protein-coding genes results **B.** Distribution of the 178 genes similar to non protein-coding elements among the three class of non coding RNAs.

3.2.1.2 Singleton genes analyses: searching for Open Reading Frame (ORFs) mistakes.

Once a gene has been sequenced it is important to determine the correct open reading frame (ORF). Every region of DNA has six possible reading frames, three in each direction. The reading frame that is used determines which amino acids will be encoded by a gene. Typically only one reading frame is used in translating a gene (in eukaryotes), and this is often the longest open reading frame. Once the open reading frame is known the DNA sequence can be translated into its corresponding amino acid sequence. An open reading frame starts with an atg (Met) in most species and ends with a stop codon (taa, tag or tga). The mis-annotation of the ORF of a protein can lead to the missing of the related paralogy relationship and so the mis-annotation of the correspondent gene as singleton. To avoid this possibility, we searched for possible erroneous ORFs among the singleton genes we detected only one mistake in the definition of the ORF for which a paralogy relationship was lost. Specifically, this is the case of two genes belonging to the MT1 family (AT1G07600 and AT5G56795): the two genes resulted as paralogs with a BLASTp analysis when the right open reading frame were considered and were afterwards removed from the list of singleton genes (Figure 3). This result reduces the list to 3408 single copy genes.

Figure 3. Wrong ORF identification of MT1B gene.

The 5'3'frame 3 is the frame as annotated in TAIR 9 database. The 3'5'frame 2 is the putative right frame. Using the first protein for BLAStp analysis no paralog genes were found. whereas using the 5'3'frame 3 we detected its paralog MT1A gene.

3.2.1.3 Singleton genes analyses: ESTs validation

The remaining 3408 singleton genes were further considered for EST validation and inter species comparative analysis (see pipeline in Figure 1). In particular, 2695 out of 3408 singletons have significant sequence similarity according to the Blastn analysis (transcript sequences versus ESTs database, E-value threshold $E \leq 10^{-5}$). The remaining 713 were classified as not confirmed by ESTs: for 695 out of 713 were discarded by the BLASTn E-value threshold of $E \leq 10^{-5}$, 24 out of 713 singletons don't show any trace of similarity with EST sequences.

Surprisingly, we found a high number of alignments in which the ESTs were longer than the aligned transcripts. We calculated the parameter delta (Δ) as the length difference between each singleton's transcript and the matched EST: genes for which transcripts were smaller than the matched EST with a $\Delta \leq 20$ nucleotides were considered as confirmed by ESTs, as this difference can be due to contaminations or EST regions not yet trimmed. The distribution of the transcripts according to the delta is reported in Figure 6.

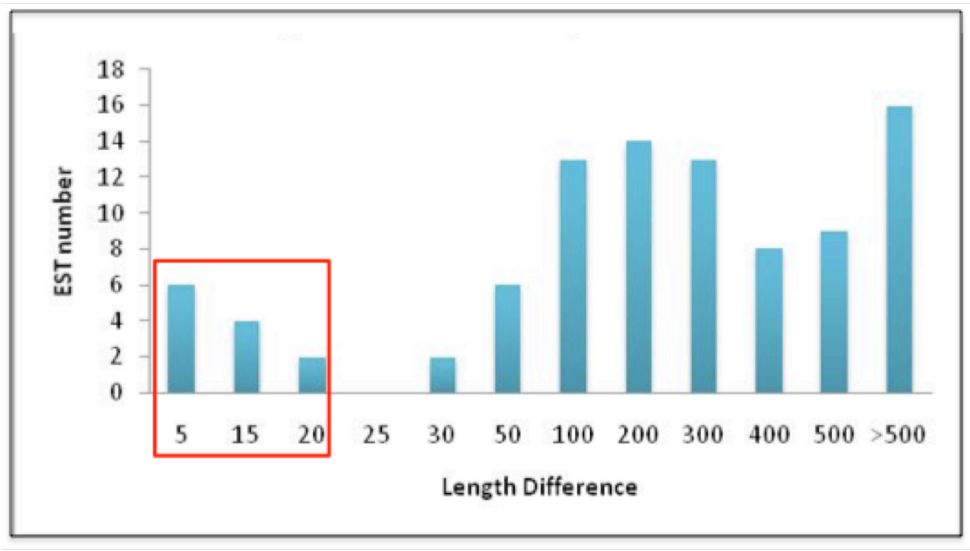


Figure 6. Lenght difference (Δ) among gene transcripts and aligned ESTs.

This histogram shows the distribution of the lenght differences in nt (x axis) among the gene transcripts and the aligned ESTs (y axis). We considered as true only those alignments with a $\Delta \leq 20$.

The 100 out of 2695 genes with $\Delta \geq 20$ nucleotides were considered as not confirmed by ESTs (Supplementary materials). Among the 272 genes filtered out by our criteria, 100 genes were discarded because of a $\Delta > 20$ nucleotides. Among the latter, only 4 are alternative spliced genes while the remaining ones have only one isoform. This means that the observed matches with longer ESTs are significant, suggesting possible mis-annotations of the gene structure annotation.

The remaining 2595 single copy genes for which we found significant sequence similarities among their transcripts and ESTs and with $\Delta \leq 20$, where further investigated as described in detail in material and methods. In brief: i) we established that the match length versus the transcript length ratio had to be greater than or equal to 60%; ii) as second threshold considering the plots shown in Figure 7, we considered that the number of identities versus the transcript or EST length ratio had to be greater than or equal to 90%.

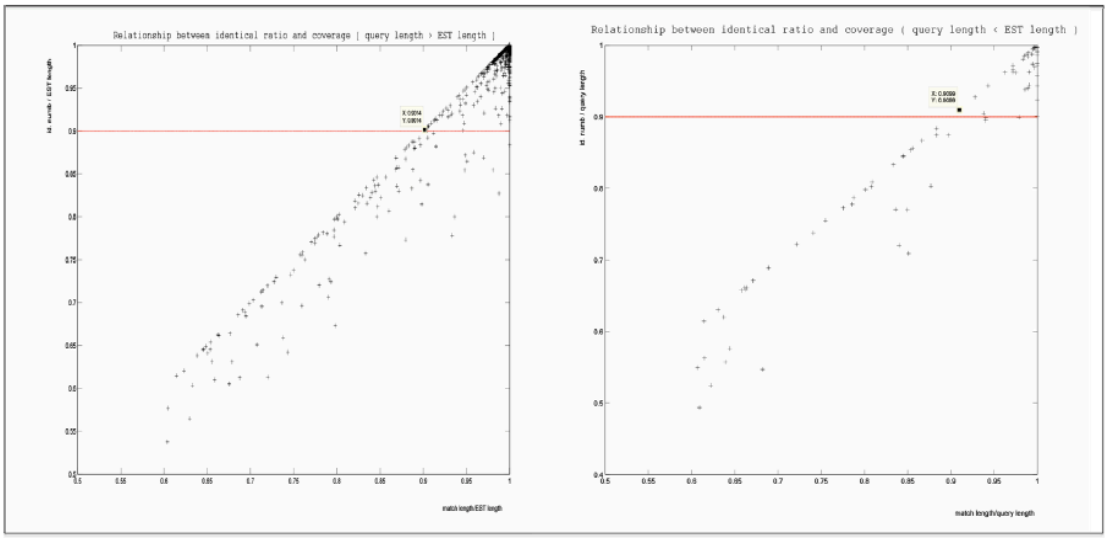


Figure 7. Validation by ESTs threshold.

When gene transcripts are longer than the aligned ESTs, we plotted the number of identities versus the EST length ratio (y axis) against the match length versus EST length ratio. The red line indicates the threshold setting (90%). The plot on the right is referred to the alignment for which the transcript is shorter than the aligned ESTs. In this case we compared the number of identities versus the transcript length ratio (y axis) against the match length versus the transcript length ratio. Also in this case, the threshold is indicated by the red line.

The application of the thresholds listed above, allowed to further classify the 2695 genes with statistical significant alignments with ESTs:

- 2414 out of 2595 genes (transcripts longer than the aligned ESTs) were classified as confirmed by ESTs;
- 9 out of 2595 genes (transcripts shorter than the aligned ESTs, with a $\Delta \leq 20$) were also classified as confirmed by ESTs;
- 172 out of 2595 genes were considered not confirmed by ESTs since the correspondent alignments don't satisfy our threshold (the number of identity versus transcript or EST length ratio greater than or equal to 90%).

These results are summarized in Table 1.

Table 1. ESTs analysis	
Genes	**Classification**
689	**Not confirmed by ESTs (excluded by E-value threshold)**
24	**No ESTs trace**
100	**Not confirmed by ESTs (ESTs longer than transcript Δ≥20)**
172	**Not confirmed by ESTs (Our threshold not satisfied)**
2414	**Confirmed by ESTs**
9	**Confirmed by ESTs (ESTs longer than transcript BUT Δ≤20)**

Table 1. EST analysis.

The genes not confirmed by ESTs are depicted in red, while the confirmed ones are reported in black.

We also checked the number of exons, the protein length (aa) and the gene length for the final list of 985 singletons not confirmed by ESTs (Figure 9). The protein products of the majority of these genes are short (about 50 aa) as well as the gene length. Moreover, they usually are miss-annotated in term of UTRs and exons coordinates. These results suggest that most of these singleton genes, which are not confirmed by ESTs, may be wrongly annotated as protein coding genes. Of course this information is important when we used the results of our analysis to investigate on the evolutionary significance of a huge number of singleton genes in a highly duplicated genome: taking into account that some of them are not protein coding genes is indeed a relevant issue.

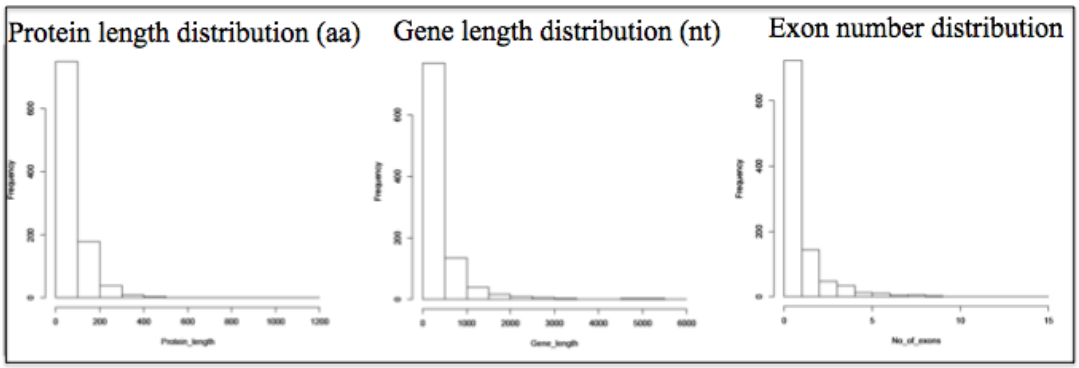


Figure 9. Structural analyses of singleton genes not confirmed by ESTs.

Protein lenght and gene lenght distributions of singleton genes not confirmed by ESTs are shown in the left and middle panels respectively. The panel on the right depicts the number of exons of the singletons not confirmed by ESTs.

3.2.1.4 Singleton genes analyses: comparative analysis

We also searched for orthologs of the singleton genes in other plant species: *O. sativa*, *P. trichocarpa*, *S. bicolor*, *V. vinifera*. We found that 1433 *A. thaliana* genes do not share orthology relationships with the considered species. Among them, 451 are confirmed by ESTs, whereas 985 are not confirmed by ESTs. Apparently, therefore, all the singleton genes not confirmed by ESTs are present only in *A.thaliana*. Among the singletons confirmed by ESTs a subset of 1450 genes have orthologs in common with all the four considered species.

Table 2. Comparative analysis

Genes	Classification
1433	No orthologs
1450	Orthologs in all the considered species
525	Orthologs in at least one species

Table 2. Comparative analysis.

The table summarize the results of the comparative analysis. We divided the genes in three classes: Genes without orthologs, genes with orthologs in all the considered species (*O.sativa*, *S.bicolor*, *P. trichocarpa*, *V.vinifera*) and genes with at least oe ortholog in one of the considered species.

3.3 Singleton genes chromosome distribution

In the previous chapter we identified a huge number of networks (1215 and 1317 respectively with E-value cut-offs $E \leq 10^{-5}$ and $E \leq 10^{-10}$) made by only two genes, in

other words, genes with only one paralogy relationship. These genes are representative of the 10% of the entire *Arabidopsis* proteome. The identification of these genes, together with the presence of singletons which correspond to the 20% of the protein-coding genes, represent an intriguing aspect in an ancient polyploid genome. In fact, from these results we assessed that about one quarter of the protein-coding genes has at most one paralogy relationship. Since we didn't expect such a high number of genes belonging to these classes in a so highly duplicated genome, we decided to visualize the singleton distribution together with the two-genes networks along the Arabidopsis chromosomes (Figure 3, Chapter 2), adding external tick marks to the Circos image in Figure 10, trying to highlight specific distributions in case. In particular, the violet ticks (D) represent the 2423 singletons confirmed by ESTs, whereas the outermost red circle depicts the 985 genes not confirmed by ESTs and without orthologs (F). The innermost green circle (E) shows a subset of the genes depicted in circle D (2423) since they include those genes confirmed by ESTs and with at least one ortholog in the considered species (1972 genes).

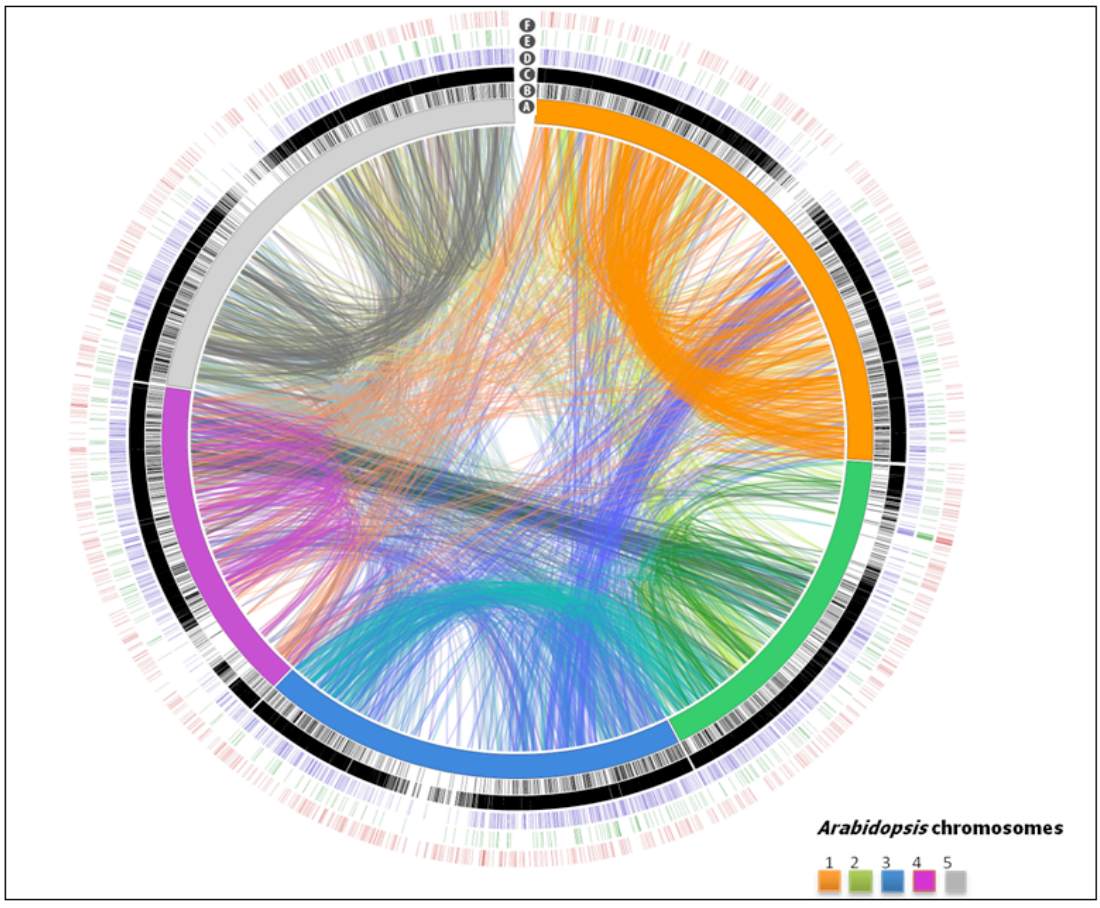


Figure 10. Distribution of paralog pairs in Arabidopsis thaliana. Singletons and two gene networks are shown.

A. The Arabidopsis thaliana chromosomes (solid colours). **B-C.** Concentric circles show gene distribution. The lines between the chromosomes indicate links between two genes (i.e. paralogs pairs). **B.** Genes involved in networks of two genes and their links are shown. **C.** All protein-coding genes except genes in the circle B and singleton genes. **D.** 2423 singleton genes confirmed by ESTs. **E.** 1972 out of 2423 singleton genes with at least one ortholog in the species considered for the comparative analysis. **F.** 985 singleton genes either not confirmed by ESTs or without orthologs.

It's worth to note that, as well as the genes involved in only one paralogy relationship, the 3802 single copy genes (circles E+F in Figure 10) are equally distributed along the five chromosomes, arising questions on the possible evolutionary mechanisms that could explain their presence in a highly duplicated genome. The massive presence of these single copy genes and their widespread distribution in the A.thaliana genome, open novel questions on the possible gene loss or gene gain mechanisms after polyploidization events.

In conclusion, the number of genes selected at each step of the singletons analyses are summarized in Figure 11.

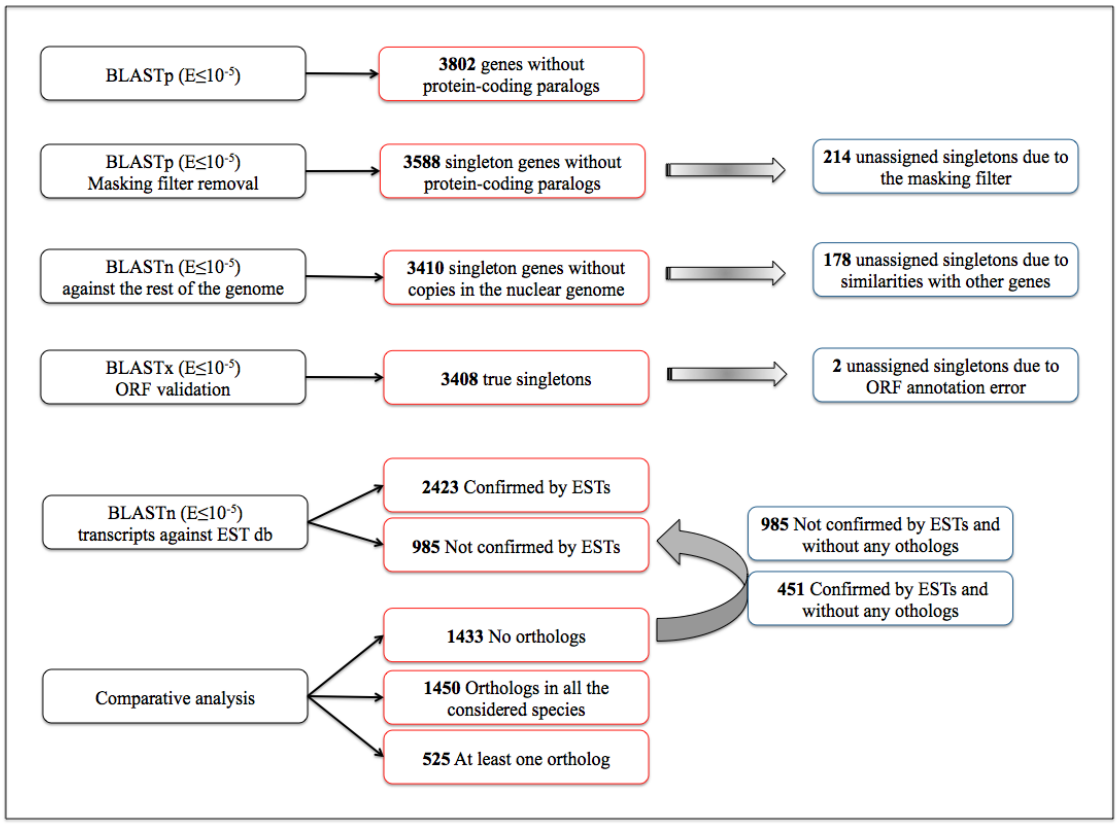


Figure 11. Singleton analyses results.

Starting from 3802 genes not having paralogs among the protein-coding genes, obtained with a BLASTp analysis using a stringent E-value cut-off ($E \leq 10^{-3}$), 214 genes were classified as "unassigned singleton genes due to the masking filter", since they are associated to other genes when considering the low-complexity regions. A Blastn analysis against non protein-coding genes allowed to exclude from the singleton's list 178 genes hence classified as "Unassigned singletons due to similarities with other genes". The Blastx analysis allowed the exclusion of other 2 genes, then indicated as "Unassigned singletons due to ORF annotation error", obtaining a final, more reliable list of 3408 single copy genes, indicated as "true singletons". The "true singletons" were divided into confirmed and not confirmed by ESTs. The comparative analysis allowed to further divide the singletons in those with at least one ortholog in the considered species, with orthologs in all the considered species and without orthologs. Merging the results from the last two analyses, the list of singletons was conclusively divided into: singletons not confirmed and without othologs, singletons confirmed and without othologs.

3.4 *Arabidopsis thaliana* Paralogy Networks Browser: singleton genes description

Singleton information was included in the genome-wide web resource presented in the Chapter 2 (http://biosrv.cab.unina.it/athparalogs/main/index). In addition to paralogies and networks information, the website allows to query the database and detect singleton genes among the whole A. thaliana collection. Moreover, for each single-copy gene, all the details about singletons are provided (i.e. confirmed/not confirmed by ESTs and orthology realationships, as reported in Figure 11). Together with the singleton and duplicated genes, the database stores also the information about the "ambiguous" genes, i.e. genes excluded either from the newtork analyses or from the singleton ones for specific settings within the pipeline. Ambiguous genes nomenclature is listed in the Supplementary materials.

3.5 Conclusions and discussions

Relatively few single copy nuclear genes (in the context of the entire genome) have been well studied in flowering plants []. Given the amount of duplications present in flowering plant genomes and their evolutionary history, orthologous sequences that are only separated by speciation events and have not been duplicated since the most recent common ancestor can be considered to be rare, and the number of genes that can be considered orthologous decreases dramatically as we compare increasingly distant lineages. The work described in this chapter allowed the identification of 3408 singleton genes, i.e. genes not having any sequence similarity with the entire nuclear genome.

3.5.1 Identification of "true" singleton genes.

The implemented pipeline, based on a protein sequence similarity search with a very loose cutoff ($E \leq 10^{-3}$), permits to first identify 3802 singletons not having copies among the protein-coding genes. Then, we also excluded from the list 178 out of 3802 genes since they showed similarities with non protein-coding genes (81 transposons, 80 pseudogenes and 41 genes coding for other classes of RNAs).

Among the “true singletons”, 1975 have at least one ortholog in the other plants considered, and 1450 out of them are shared between all the 4 species. Though the number here reported is not exactly the one reported in other literature ref, it has been already argued the evolutionary issue concerning the presence of low or single copy genes per taxa [citare lavori 4-6 in Duarte e DUARTE] as well as the structure and functional properties of these conserved single copy genes among species [Duarte]. However, we raise the issue that the annotation of part of these singletons would need further confirmation, since the 985 out of 1433 genes not having orthologs in the plant species considered, also do not have clear confirmation based on EST evidence. Unfortunately, 24% of them is neither represented within the affymetrix chip (data not shown) therefore there is not useful evidence for their reliable annotation. Therefore, the complete collection of singletons classified according to the features established during the presented analysis has been made public to permit comparisons and confirmation of the issues here highlighted. Furthermore, previous analysis of singletons in plant species, *A.thaliana* included (TAIR 9 release), highlighted that the amount of singleton genes was around 5000 genes (Duarte 2010). Our pipeline, which takes into account different computational issues, such as the use of the masking filter for the sequence alignment, allows to reduce the dataset of true singletons to 3408 genes. The refinement of the data here presented is essential to draw more reliable conclusions both on the quality of the annotation and on the issues related to singleton presence in a widely duplicated genome, such as diploidization and fractionation mechanisms ref.

3.5.2 Highlights for the need of a more accurated annotation process for the model plant species.

It is reasonable to imagine that not all the genes annotated by the TAIR as protein coding should emerge from our analysis as confirmed by EST, as many genes may have not this confirmation. Indeed, however, this is not the case for 985 protein-coding genes which are marked, by our analysis, as not confirmed by ESTs. In particular, trace of similarity was found for 24 out of 985 genes, while 689 genes were considered not confirmed by ESTs since the statistical significance of the alignment is above the considered threshold. Very interesting is the evidence of 100 genes for

which the aligned ESTs are much longer than the annotated region, which confirms the mis-annotation of the gene loci today identified; moreover, there is no evidence for annotated alternative transcripts in the same loci. Further support to doubts concerning the annotation of these genes is gained from the comparative analysis: in fact all the not confirmed by ESTs genes are present only in A. thaliana, sharing no ortholog with other species. On the contrary, among the 2423 genes confirmed by ESTs, 1972 genes have at least one ortholog in the considered species whereas 451 are present only in *Arabidopsis thaliana*. We further classified the gene ontologies specific of singletons confirmed by ESTs and have no orthologs in the other plants here considered. Since these genes have no confirmation from ESTs and result unique representatives within a highly duplicated genome and among related species, they highlight the need of further investigations since any possible fault in their annotation my affect the exploitation of A. thaliana in terms of comparative genomics.

The absence of ESTs or not suitable gene structure related to the present ones supporting the annotation of the genes as protein coding ones, pushes versus a necessary review and validation of the annotations themselves and maybe highlights the need of a more accurate annotation process.

3.5.3 Singleton genes within the *Arabidopsis thaliana* paralogy network browser.

Since the current analysis may suggest further intriguing issues concerning the gene organization of A. thaliana, supporting genome annotation and evolutionary analyses, we added our results into the web accessible database, accessible online at the address http://biosrv.cab.unina.it/athparalogs/main/index (Chapter2). As described in the previous chapter, this resource allows to query the *A.thaliana* genome by gene, or gene family, or keyword (i.e transcript name). The results, organized in a table, provide information about the singleton genes and their classification. This permits to further exploit the full use of *A.thaliana* as reference, with more informative detailed included for all the mRNA encoded genes included in its annotation. The web based resource permits the user to distinguish among protein-coding genes which are in single copy in the genome, that have been confirmed by ESTs and have orthologs in

other plant species, from genes similar to other classes of genes or with an ambiguous annotation (i.e. genes not confirmed by ESTs and without orthologs).

Moreover, the entire collection of our results represents a resource useful to futher investigate evolutionary issues that could unravel the complex and mysterious history of Arabidopsis thaliana genome.

3.6 Material and Methods

3.6.1 Data retrieval

The *Arabidopsis thaliana* genomes (release TAIR9, June 2009 and TAIR 10, November 2010), were downloaded from the TAIR website (http://www.*Arabidopsis*.org/). TAIR9 consists of 27,379 protein-coding genes, 4827 pseudogenes or transposable elements and 1312 ncRNAs. TAIR10 consists of 27,416 protein coding genes, 4827 pseudogenes or transposable element genes and 1359 ncRNAs. We removed from protein coding genes the mitochondrial and chloroplastic ones, reducing the TAIR9 and the TAIR10 datasets to 27,169 and to 27206 genes, respectively.

3.6.2 The pipeline

In the first step of the pipeline, an all-against-all protein sequence similarity search was performed using the BLASTp program [Altschul S.F., 1997]. To obtain a reliable determination of protein paralogy relationships we considered the following criteria:

the E-value cut-off. Since we are aimed to find only genes present in single copy among the Arabidopsis protome complement, the BLASTp analysis was based on a no stringent E-value cut-off ($E \leq 10^{-3}$). This allows to avoid the possibility to consider as singleton a genes too much diverged from its copy.

the twilight zone. The aligned sequences were further filtered with the Rost's formula, [Rost B, 1999], to determine whether two proteins in a genome are paralogs when the similarity between them is in the so-called twilight zone (20-30% of identity/length ratio). The Rost's formula was applied with the

cut-off threshold defined by Li et al., 2001. These genes were removed from the singleton analyses due to the intrinsic ambiguity of their paralogy relationship.

the low-complexity BLAST filter. By default BLAST applies a masking of low-complexity regions in the query sequence [Wootton and Federhen, 1996]. Due to their unusual composition and repetitive nature [Golding, G.B. 1999], the presence of low complexity regions can result in biologically meaningless high scoring hits [Sharon I, 2005]. However, sequence similarities can often be missed due to the masking of low complexity regions, particularly in the case of smaller sequences. To avoid this problem we performed the BLASTp analysis twice: in the first analysis the default parameters were used utilizing either the more stringent cut-off or the less stringent one. A second BLASTp analysis was performed using the E-value cut-off E≤10-5 and without the masking filter, in order to recover genes that could have been masked hiding similarities to other genes. The sequences showing similarity when unmasked were neither included in the paralogs class, nor considered singletons.

3.6.3 Singleton genes analyses: protein-coding genes versus other regions.

Proteins without similarity versus the *A. thaliana* protein collection (based on the unmasked BLASTp, cut-off $E \leq 10^{-5}$) were further investigated to discard any other similarity with other regions of the entire genome. The full length genes were thus compared with both non protein-coding genes and inter-genic regions (based on a BLASTn analysis, E-value threshold $E \leq 10^{-5}$)

Non protein-coding genes (pseudogenes, snRNA, snoRNA, miRNA, tRNA, rRNA, other RNA and transposable elements) were classified based on the locus type classification assigned by the TAIR. Further support in the pseudogene classification was achieved considering the file "TAIR9 functional annotation" from the TAIR website (http://www.Arabidopsis.org/), where all the relationships among pseudogenes and protein-coding genes reported in Zhang et al. 2006 are also included. The resulting alignments versus this collection were manually curated. We discarded the matches involving exclusively the intronic regions of the query, and the

matches corresponding to regions of the query in which known non protein-coding genes are embedded. We accepted as positive results the matches involving exonic regions longer than the half of the length of the query.

The protein coding genes resulting as not having any similarity versus other genes, were compared with the intergenic regions (with a BLASTn analysis, E-value$\leq 10^{-5}$) to exclude any significant sequence similarity within the whole genome.

3.6.4 Searching for ORF annotation errors.

To check for possible annotation mistakes due to error in the ORF definition, the *Arabidopsis* transcript sequences were aligned versus the protein collection with a BLASTx analysis (E-value$\leq 10^{-5}$). The ExPASy Translate tool [Gasteiger E., et al., 2003] was also considered to further check the best ORF for a transcript. Genes not matching any region of the A. genome were annotated as singletons.

3.6.5 Validation of singleton genes with Expressed Sequences Tags evidence.

The complete set of *Arabidopsis thaliana* EST sequences was downloaded from the GenBank release of April 8, 2010. Transcript sequences from singleton genes were aligned to EST sequences (with a BLASTn, E-value$\leq 10^{-5}$).

We defined a delta (Δ) corresponding to the EST length minus the transcript length. Then, we divided the alignments into two classes, classifying transcripts smaller ($\Delta \leq$20nts) or longer ($\Delta >$20nts) than the matching EST.

The two classes were independently analyzed:

1) transcripts smaller or equal than the matching EST: among the alignments where the match length versus the transcript length ratio was greater than or equal to 60%, we considered significant only those where the ratio between the number of identityies and the transcript length is greater than or equal to 90% (Suppl. Fig 1B).

2) Transcripts longer than the matching ESTs: among the alignments where the match length versus the EST length ratio was greater than or equal to 60%, we considered significant only those where the ratio between the number of identities and the EST length is greater than or equal to 90% (Suppl. Fig 1B).

The occurrence of significant matches permitted to classify the singletons as confirmed by EST evidence.

3.6.6 Searching for orthologs.

Singleton genes were used as input to search the Ensembl Plants database (Plant Mart Release 5) [http://plants.ensembl.org/biomart] for the detection of orthologs genes using the BioMart multi species comparisons tool [Smedley D, et al., 2009]. We considered only the species available within the BioMart tool*: Oryza sativa, Sorghum bicolor, Vitis vinifera and Populus tricocharpa.*

Chapter 4

Arabidopsis thaliana transcription factors: classification and investigation of dosage sensitive genes in a highly duplicated genome

4.1 Introduction

One single transcription factor may control the expression of several target genes and one target gene can be regulated by many different transcription factors in the frame of even very complex regulatory networks. Duplication has played a key role in the evolution of interactions in the regulatory networks []. In fact, both transcription factors and the target genes have been extensively duplicated. Following duplication, interactions may be inherited from the ancestor to the duplicate or new interactions may be gained through divergence. According to a study by Teichmann and Babu (2004), approximately half of the interactions have been gained through sequence divergence after duplication, whereas one-third of the interactions have been inherited from the ancestral gene. Only a minority (10%) of the interactions between transcription factors and target genes consist of genes that lack homologs []. This introduction will give an overview of the evolution of transcription factor genes and their classification within the Plant kingdom. Moreover, an elucidation about the reasons why the study of transcription factors is still an intriguing issue in a reference genome such as the Arabidopsis one will be provided using a well-known evolutionary hypothesis.

4.1.1 Evolution of transcription factor genes

Alterations in the activity and regulatory specificity of TFs are now established as major sources for species diversity and evolutionary adaptation [6]. Indeed increased complexity of the regulatory system appears to have been a principal requirement for the emergence of metazoan life. The distribution of TF families across the extant eukaryotic lineages presumably arose through a series of varied evolutionary events, including the de novo appearance of functional peptides and protein domains, adaptation of ancestral proteins into novel TFs, family expansions in the form of duplication and divergence, lineage-specific losses and the acquisition of novel TFs through horizontal transfer. The evolutionary history of protein families can be inferred by considering their phylogenetic distribution, providing an estimate of when (and from whence) the various families arose. A summary of the origins and

expansions of the major eukaryotic TF families is depicted in Figure 1.

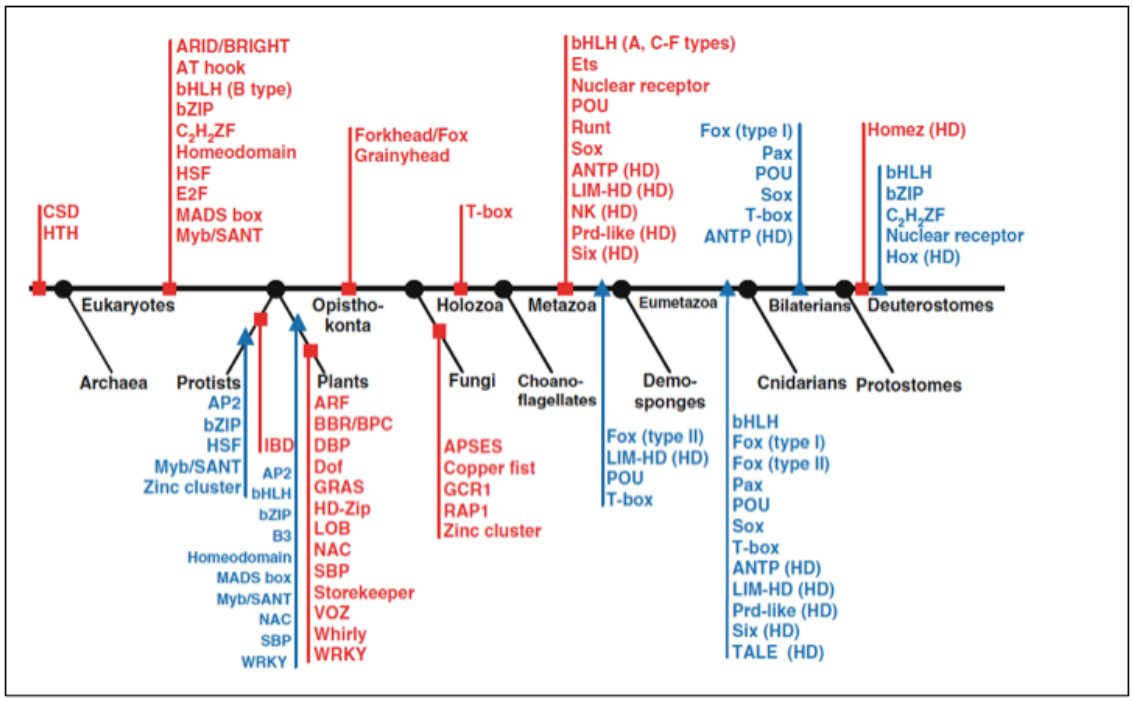


Figure 10. Evolutionary timeline of transcription factor families.

Estimated timing of the appearance (red) of major TF families and subfamilies, as well as large-scale family expansions (blue). Homeodomain subfamilies are indicated as "HD". Brief definition of select clades: Opisthokonta: fungi and metazoa; Holozoa: animals and choanoflagellates; Choanoflagellates: free-living unicellular and colonial flagellate organisms; Eumetazoa: animals, excluding sponges and some other simple animals; Demosponges: group containing the majority of sponges; Bilaterians: all animals with bilaterial symmetry; Cnidarians: aquatic jelly-like organisms, including jelly fish, corals, and sea anemones; Deuterostomes: vertebrates, echinoderms (e.g. sea stars and sea urchins), tunicates (e.g. sea squirts) and others; Protostomes: insects, crustaceans, nematodes, flatworms, mollusks, brachiopods, and others. Adapted from Harris [Harris et al., 2011].

4.1.2 Gene duplication of dosage sensitive genes: an intriguing issue in *A.thaliana*

The Gene Balance Hypothesis predicts that an imbalance in the concentration of protein subunits in a macromolecular complex or between proteins with opposing functions in a transcription or signaling network may either lead to decreased fitness or lethality (Birchler and Newton 1981; Birchler et al. 2001; Veitia 2002; Papp et al.

2003; Veitia 2005; Veitia et al. 2008). Maintaining proper protein and transcriptional balance is vital to sustain normal functions. For instance, an imbalance in a highly connected portion of a regulatory network, likely would result in largely negative pleiotropic effects. For example, a modification in the relative abundance of subunits in a transcription factor complex may alter the assembled complex and the expression of target genes (Birchler et al. 2001). The Gene Balance Hypothesis, supported by analyzed genomes across eukaryotic lineages, provides the basis for understanding duplicate retention following gene and genome duplication. For instance, dosage-sensitive genes must be retained as duplicates following WGD to maintain proper balance of protein and transcriptional networks. However, following a smaller scale duplication (e.g. local and tandem duplicates, segmental duplicates, aneuploidy), duplicates of dosage-sensitive genes will tend to be eliminated to maintain proper balances. Whole genome duplications differ from smaller scale duplications in that the first ones increase the dosage of all genes simultaneously. Thus, organisms experiencing WGD like *Arabidopsis thaliana,* immediately maintain proper balance in both signaling and transcription networks as well as stoichiometric balance in macromolecular complexes. During diploidization, the spectrum of remaining duplicates would be expected to be random if gene loss is neutral. However, comparative genomic studies have revealed that gene loss is not random, which begs the question as to whether selection operates to either retain gene duplicates, return genes to single copy, or both. Interestingly, some functional gene categories, including subunits of protein complexes such as transcription factors and ribosomal proteins, are significantly over-retained in duplicate and have resisted loss during the diploidization process in the *Arabidopsis* (Maere et al. 2005; Freeling 2008), Paramecium (Aury et al. 2006), vertebrate (Blomme et al. 2006), and yeast (Papp et al. 2003) genomes. In other words, the co-retention of interacting dosage-sensitive genes, following a WGD and during diploidization is necessary to maintain proper balance of dosage-sensitive complexes and networks.

In this frame, the analysis of transcription factor genes in terms of duplicated genes in an ancient polyploid organism, which underwent dipolidization and reshuffling of the gene content, represents definitely an interesting issue. Moreover, understanding how transcription factor gene families are organized in the model species *A.thaliana* can help the annotation and structural and evolutionary investigations of transcription factors in other plant species.

4.1.3 Plant transcription factor genes according to the literature and publicly available databases

In general, between 3 and 6% of green plant (*Viridiplantae*) genes encode TFs. Interestingly, many plant-specific DNA-binding domains (DBDs) bind DNA utilizing β-sheets, in contrast to other eukaryotes and prokaryotes, which largely bind DNA utilizing α-helices [67]. Major eukaryotic TF families present in plants include Myb/SANT, bHLH, bZIP, Plant β-Scaffold Transcription Factors MADS box proteins are found in a variety of organisms, but are most prominent in the genomes of flowering plants, where they have undergone multiple expansions stemming from whole genome duplication events [68]. MADS box TFs control all major aspects of the life of green plants, and the timing of the expansion of this family suggests that it might have played a key role in the evolution of flowering plants [69]. B3 domains are present in three families of plant TFs: Auxin response factors (ARFs), VP1, and RAV-like AP2 TFs (which also have AP2 domains). The B3 domain consists of a β-sheet and two α-helices situated between β-strands, with loop residues predicted to make deep contacts with the major groove of DNA [72]. Although believed to be plant-specific, B3 domains share structural similarities with the Escherichia coli EcoRII restriction enzyme, suggesting a possible horizontal transfer event between a eubacterial ancestor to an ancestral plant (or vice versa) [67]. Members of the Whirly TF family are found throughout the plant kingdom, and play roles in the regulation of genes involved in defense response [73]. Whirly TFs bind DNA as tetramers, with each unit consisting of two anti-parallel β-sheets packed perpendicularly against each other [74]. The four resulting blade-like extensions adopt a striking “whirligig-like” appearance, providing the namesake for this family of TFs. Whirly protein complexes bind single stranded DNA, in contrast to the majority of eukaryotic TFs [74]. Plant Zinc-Coordinating Transcription Factors WRKY TFs, are zinc-coordinating proteins that adopt a β-sheet fold [75]. Although initially believed to be plant-specific, WRKY proteins have a similar fold to metazoan GCM family TFs and fungal Rcs1p and Rbf1p proteins [77], suggesting a more ancient evolutionary origin. SBP family TFs are involved in a variety of developmental processes, including flower development in particular [79]. SBP DBDs include a pair of zinc binding sites consisting of eight

residues in a novel C3H, C2HC or C6HC configuration, with the first four residues coordinating one zinc atom, and the last four coordinating the other [81] (there is, as yet, no structure of an SBP domain in complex with DNA). The Dof family is composed of a diverse range of proteins that bind DNA utilizing a C2C2 zinc finger [82]. Like many other zinc fingers, the Dof domain can also function in the mediation of protein-protein interactions, often with members of the bZIP family [83]. Dof TFs can be classified into six distinct subfamilies, each with a unique domain architecture [84]. Additional Plant Transcription Factor Families AP2 proteins comprise the largest family of TFs that are mostly specific to plants, although members have recently been identified in a wide range of organisms, including TFs in apicomplexans (unicellular animal parasites) [86], and endonucleases in prokaryotes [87], bacteriophage [88], and yeast [89]. NAC proteins comprise the second-largest family of plant TFs, with the *Arabidopsis thaliana* genome containing over one hundred putative NAC TFs. Structurally, the central four strands of the NAC monomer are highly similar to the four-stranded β-sheet of the WRKY domain, suggesting an ancient evolutionary relationship [67, 77]. The Myb/SANT family, though structurally distinct from homeodomains, also binds DNA utilizing an HTH-based DBD. Myb/SANT proteins can contain up to three imperfect repeating DNA-binding α-helical sections. In plants, Myb/SANT TFs often interact cooperatively with members of the bHLH family, as exemplified by the regulatory control of the phenylpropanoid biosynthetic pathways [94, 95]. Myb TFs are most abundant in plants, but are prevalent across the eukaryotic kingdom (Fig. 3.2). Members of the Myb family are known to be factors in several human cancer subtypes [96].

An heterogeneous classification of plant transcription factor gene families here described is available in different plant TF databases. The latter are surely a valuable tool for plant biologists to study plant physiology, plant developmental biology and plant stress physiology. The challenge is represented by different classifications in distinct databases, due to different methods used for families annotation and/or different genome releases. The wide number of plant TF databases freely available over the web is definitely a valuable resource for plant molecular biologists, often, however, determining some confusion and certainly, even, still needing a refinement in the transcription factors annotation and classification, as well as in the classification of their relative families.

4.1.3 Aim of the chapter

The presence of widespread intra-genome duplications, together with the loss of many gene copies in the *A. thaliana* genome, really complicates the interpretation and the study of the evolution of TF gene families within this species, threatening the role of this genome as a reference in plant comparative genomics. A major limitation to studying TFs is the lack of a reliable and unique annotation, a challenge that is compounded by the presence of many dedicated databases in which specific and varied methodologies for the identification of genes within a family are considered.
In this chapter, we focused on well-known resources of TFs collections obtained from different databases. Performing an INTERPROscan analysis, we refined the annotation of these genes, classifying the latter in two categories: transcription factor genes (i.e. genes with at least one DBD) and co-regulatory genes (i.e. genes with regulatory domains which help TF activity without presenting the DBDs –hai definite prima questo acronimo). Then, we performed a deeper investigation on TFs organization within the *A. thaliana* genome, via the analysis of networks of paralogs (results described in Chapter 2). In particular, we used network and singleton data (Chapter 2 and 3) to investigate on gene families evolution, underlying structural relationships between TFs belonging to different families and/or between TFs and non-TF genes. Moreover our work provides support to the classification of TFs and coregulators in *Arabidopsis thaliana* and represents a step forward to understand TF family organization and evolution in this species.

4.2 Results

4.2.1 Overview and integration of the major databases

The first step of our analysis consists in the integration of data from four different plant transcription factor databases: PlantTFdb [], PlnTFdb [], AtTFdb [], DATF []. Among them, AtTFdb and DATF *A.thaliana* specific, while the first two store information about multiple plant species, among which *Arabidopsis*. In Table 1, the summary of the information concerning *Arabidopsis* transcription factors in each

database is shown. As described before, transcription factors are always grouped into different families based on their DNA binding domains. However, there are profound differences among the different databases both in terms of transcription factors collections and number of families in which the latter are organized. Moreover, each database may be based on a different genome releases.

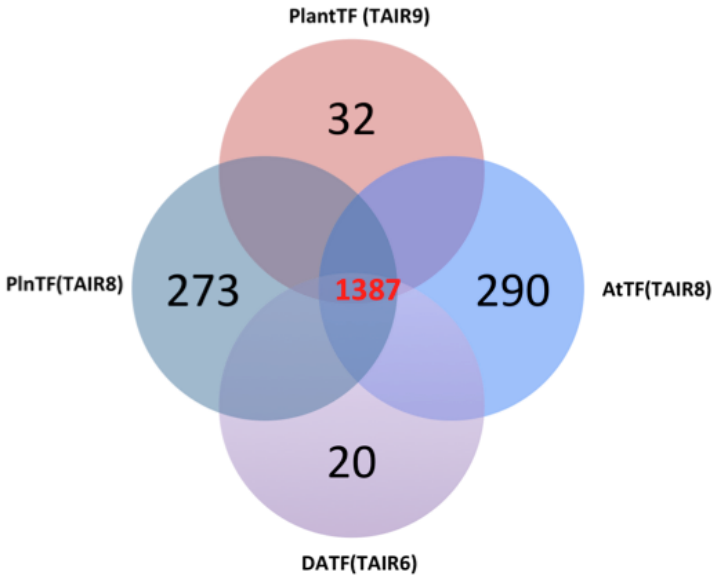
Table 1. Publicly available *Arabidposis* transcription factor databases			
Name	**Genome release**	**Transcription factor genes**	**Transcription factor gene families**
PlantTFdb	**TAIR 9**	**2023**	**58**
Agris (AtTFdb)	**TAIR 9**	**1841**	**51**
PlnTFdb	**TAIR 8**	**2192**	**82**
DATF	**TAIR6**	**1922**	**64**

Table 1. Plant transcription factor databases.

The table shows the number of *Arabidopsis thaliana* transcription factor genes (third column) present in each of the four plant TF databases considered, as well as the number of family in which these TFs are organized (last column). Moreover, the genome release used in the different databases is depicted (second column).

We searched for genes in common among the considered databases. We found that 1387 genes are annotated as transcription factors in the four databases, while a high number of other genes are typical of each database (Figure 1A). The table in Figure 1B depicts the number of TFs in common between each database pair.

A



B

	PlantTFdb	AtTFdb	PlnTFdb	DATF
PlantTFdb	2023	1445	1627	1578
AtTFdb		1841	1485	1490
PlnTFdb			2192	1849
DATF				1922

Figure 1. Integration of four plant transcription factor databases.

A. The Venn diagram shows the number of genes (in red) annotated as TFs in the four databases considered. The number of TFs specific of each database is also indicated.

B. TFs in common between each pairs of databases.

Transcriptional regulators (or transcription cofactors) are proteins that interact with transcription factors to either activate or repress the transcription of specific genes. These proteins can be detected by the presence of distinctive domains (activator and/or inhibitory domains) and for the absence of DNA binding domains. PlnTFdb is the only database, among the ones considered in this work, which considers both TFs (transcription factors) and TRs (transcription regulators). On the other hand, PlantTFdb includes only transcription factors in its collection, as those proteins that show specific DNA binding sites and are capable of activating and/or repressing transcription, excluding transcription cofactors and chromatin related proteins like chromatin remodeling factors, histone demethylases, histone acetyltransferase. The other two databases (AtTFdb and DATF) didn't take into consideration these issues.

The effort of integrating the four databases allowed to get a novel classification:

1. Confirmed TFs: those genes classified as transcription factors in all the considered databases. Since PlantTFdb is based on the last genome release and the pipeline implemented for the classification is consistent, we consider the latter as the more reliable collection. Indeed, genes annotated in PlantTFdb and in at least another database are considered as confirmed TFs.
2. Putative TFs: those genes not annotated as transcription factors inPlantTFdb but present in at least two other databases.
3. Co-regulators: transcriptional regulator genes (annotated only in PlnTFdb)

According to the criteria listed above, we confirmed 1590 out of 2830 total genes as transcription factors, while 951 genes were classified as putative TFs. The remaining 289 genes were classified as coregulators.

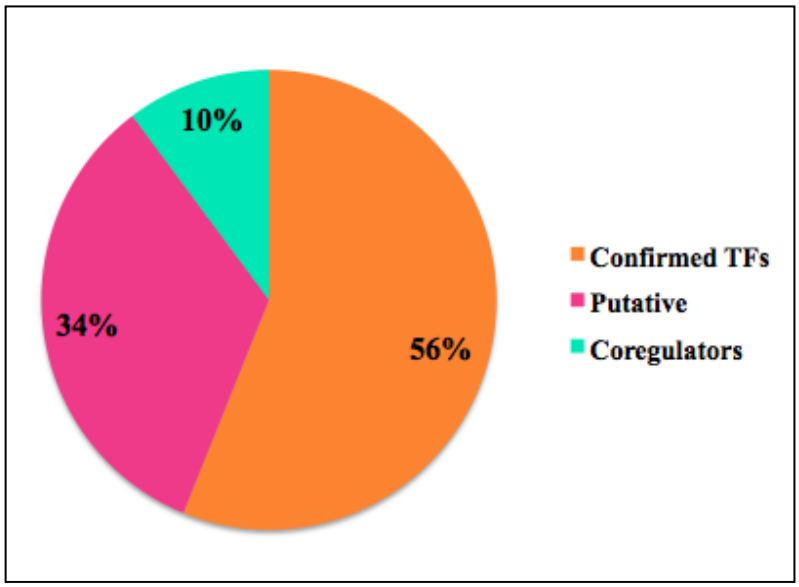


Figure 2. Preliminary classification resulting from the database integration.

The results from the integration of the different databases are shown: in orange and in pink the percentages of confirmed and putative TFs respectively, are shown. In green the percentage of coregulators is shown.

This preliminary glance makes already clear the absence of a reliable and unique classification of *Arabidopsis* TFs. This makes challenging the use of this genome as

reference for the annotation of transcription factor gene families in other plant species. In this frame, we decided to overcome the difficulties due to heterogeneous information about *A.thaliana* transcription factors, obtaining a novel classification based on the most recent genome release (TAIR10).

4.2.2 A novel classification of *A.thaliana* transcription factor genes

To identify the repertoire of transcription factors in the *Arabidopsis* genome, we used the list obtained by the integration of the available databases and we define a class of genes encoding for proteins that binds DNA in a sequence specific manner, distinguishing them from coregulators, i.e. proteins that interact with transcription factors to either activate or repress the transcription of specific genes.

First, we reduced the initial list of 2830 genes got from the integration of the databases to 2802 genes since the TAIR10 release was used. In fact, 6 genes are annotated as pseudogenes and transposons (4 and 2 out of 6 respectively), whereas 22 out of 2831 are not annotated as genes anymore in the most recent genome version.

Then, we assembled a list of DNA-binding and coregulatory domains from the InterPro database (release 17). For each entry we examined the description, the associated literature and their presence in the considered databases, to assess their sequence specific DNA-binding capabilities. This analysis resulted in an manual curated list of 150 domains (Supplementary material). We then extracted 2523 out of 2802 genes showing a significant match with the selected DNA-binding domains: they were classified into transcription factor families according to the identified domains and to the other database information. The remaining 279 genes were considered as “orphans”, as they contain one or more domain(s) whose presence, or combination, according to the literature, does not allow their classification into any of the defined families. Moreover, these genes are not associated to any family in all the considered databases.

In doing so, we present a comprehensive and high quality list of 2802 TFs/coregulators in the *Arabidopsis* genome distributed in 99 gene families plus 279 genes falling in the category of orphan genes.

We further divided these genes in three classes:

- TFs class A: genes having at least one DNA-binding domain (DBD).
- TFs class B: genes having either the DBD or a regulatory domain
- Coregulators: genes having regulatory domains.

The 47% (1316 genes) of the genes are classified TFs class A, while the 21% (592 genes) belong to the second class (TFs class B). Genes belonging to these two classes are considered as TFs since the presence of at least one DNA-binding domain, obtaining a final list of 1908 transcription factors. Coregulators represent the remaining 32% (894 genes).

Our analysis allowed the accurate classification of the putative TFs obtained by the integration of the databases: 283 out of 951 putative TFs were classified as transcription factor genes, whereas 641 of them belong to the coregulator class in the new classification. As an additional evidence of the ambiguous nature of the putative genes, 27 out of 28 genes are not anymore annotated as protein-coding ones in the new genome release (TAIR10) and were excluded from the final list. On the other hand, 82 genes classified as coregulator in the previous databases, are in our classification considered as transcription factor genes. In contrast, 46 genes annotated as transcription factors in the other databases, belong to the coregulator class in the new classification (Table 1).

Table 2 . Old versus new classification: improved annotation of transcription factor genes.

Old classification (Databases integration)	New classification
Confirmed TFs	1543 TFs (Classes A and B) 46 Coregulators 1 excluded by new TAIR release
Coregulators	82 TFs (Classes A and B) 207 Coregulators
Putative TFs	283 TFs (Classes A and B)
	641 Coregulators
	27 excluded by new TAIR release

Table 2. Old versus new classification: improved annotation of transcription factor genes.

The table shows how the genes belonging to the different classes inferred from the integration of data from available databases are classified in the new classification resulting from our analyses.

4.3 TFs and coregulators duplicability: singleton or duplicated genes?

Exploiting the results obtained in Chapters 2 and 3, we investigated the transcription factor and coregulator genes in association to the duplication events. In particular, we investigated on the distribution of these genes among the networks of paralogs (either the more and less conservative ones) and among singleton genes (see Chapters 2 and 3 for more details).

The numbers of TF and coregulator genes in the two classes of networks are depicted in Table 2. Networks containing both TFs and coregulators are 40 and 39 using the more stringent and less stringent E-value cut-offs, respectively.

Table 3. TFs and coregulators distribution among networks of paralogs and singleton genes

Class	Networks of paralogs	Number of genes	Number of networks
Transcription factors	More conservative networks ($E \leq 10^{-10}$)	1828	153
	Less conservative networks ($E \leq 10^{-5}$)	1856	106
Coregulators	More conservative networks ($E \leq 10^{-10}$)	808	163
	Less conservative networks ($E \leq 10^{-5}$)	829	135

Table 3. TFs and coregulators distribution among paralog networks

The number of duplicated TFs and coregulators obtained using two different E-value cut-offs ($E \leq 10^{-10}$ and $E \leq 10^{-5}$) are shown together with the number of networks which they belong to.

Since our aim was the identification of reliable and consistent relationships in which these genes were involved in, we focused our attention on TF and coregulator gene families distribution among the more conservative networks, i.e. the one obtained with a more stringent E-value threshold ($E \leq 10^{-10}$).

As described in Chapter 2, the organization of duplicated genes in networks of paralogs is helpful not only to reorganize data from a complex model genome, but also to study gene families in terms of structural relationships with genes belonging to the same family and/or belonging to two or more different gene families. Moreover, networks of paralogs can be used to refine the annotation of orphan genes. Both these two applications were handled to study transcription factor and coregulator gene families in *A.thaliana*.

4.3.1 Exploiting networks of *A. thaliana* Transcription Factors

Transcription factor and coregulator genes are distributed among different conservative networks, with different sizes and complexity. We classified the 99 TF/coregulator families on the bases of their distribution among the networks of paralogs, as follows:

- Exclusive family: all the genes belonging to that family are arranged into a single network
- Distributed family: the genes belonging to that family are spread across different networks.
- Singleton family: all the genes of the family are not present in networks at least when using the more stringent E-value cut-off.

54 out of 99 families are included in the first class while 38 families were classified as distributed ones. The remaining 4 families were considered single, since all the genes included in the families are indeed not contained in networks (Figure 4).

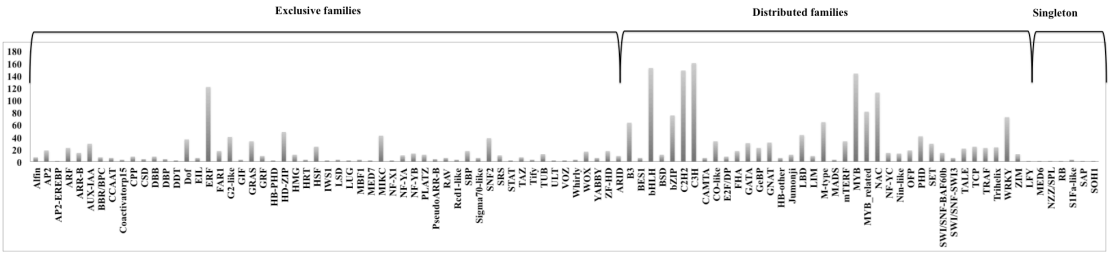


Figure 4. Families classification based on their distribution among networks of paralogs.

The histogram shows the number of genes (y axis) in each family (x axis). The bars above the histogram depict the type of family according to our classification.

Networks, in turn, were classified as follow:

- Exclusive TF/coregulator network: the network contains one exclusive family.
- Exclusive-distributed TF/coregulator network: the network contains only the genes of a distributed family (this means that some members of the family are contained also in other networks)
- Mixed TF/coregulator: the network contains multiple TF and or coregulator families, which may be either exclusive or distributed.

- Inclusive: the network contains TF/coregulator families (absolute or distributed) as well as other genes do not belonging to non-TF or coregulator genes.

We found 156 out of 276 networks containing only genes belonging to one TF/coregulator family: 44 out of 133 are exclusive networks, since all the components of the gene family are herein contained, while 112 out of 133 are exclusive-distributed ones, since other genes belonging to the same family are also distributed in other networks. On the other hand, we found 40 out of 276 mixed TF/coregulator networks. Finally 80 out of 276 networks were classified as inclusive ones, since they contain also genes not annotated as TFs and or coregulators (Table 4). It's worth to note that either mixed or inclusive networks can contain the entire component of a TF/coregulator family but at the same time include also other genes

Table 4. Networks classification	
Type	**Number of networks**
Complete Exclusive TF/coregulator	**44**
Split Exclusive TF/coregulator	**112**
Mixed TF/coregulator	**40**
Inclusive	**80**

Table 4. Networks classification.

The types of networks of paralogs assigned on the bases of the gene families herein contained are indicated in the first column. The number of networks for each class is reported in the second column.

We summarized the information collected in Figure 5. Red nodes represent mixed and inclusive networks, while the green ones depict exclusive and exclusive-distributed networks. Blue and orange nodes are TF and coregulator families, respectively.

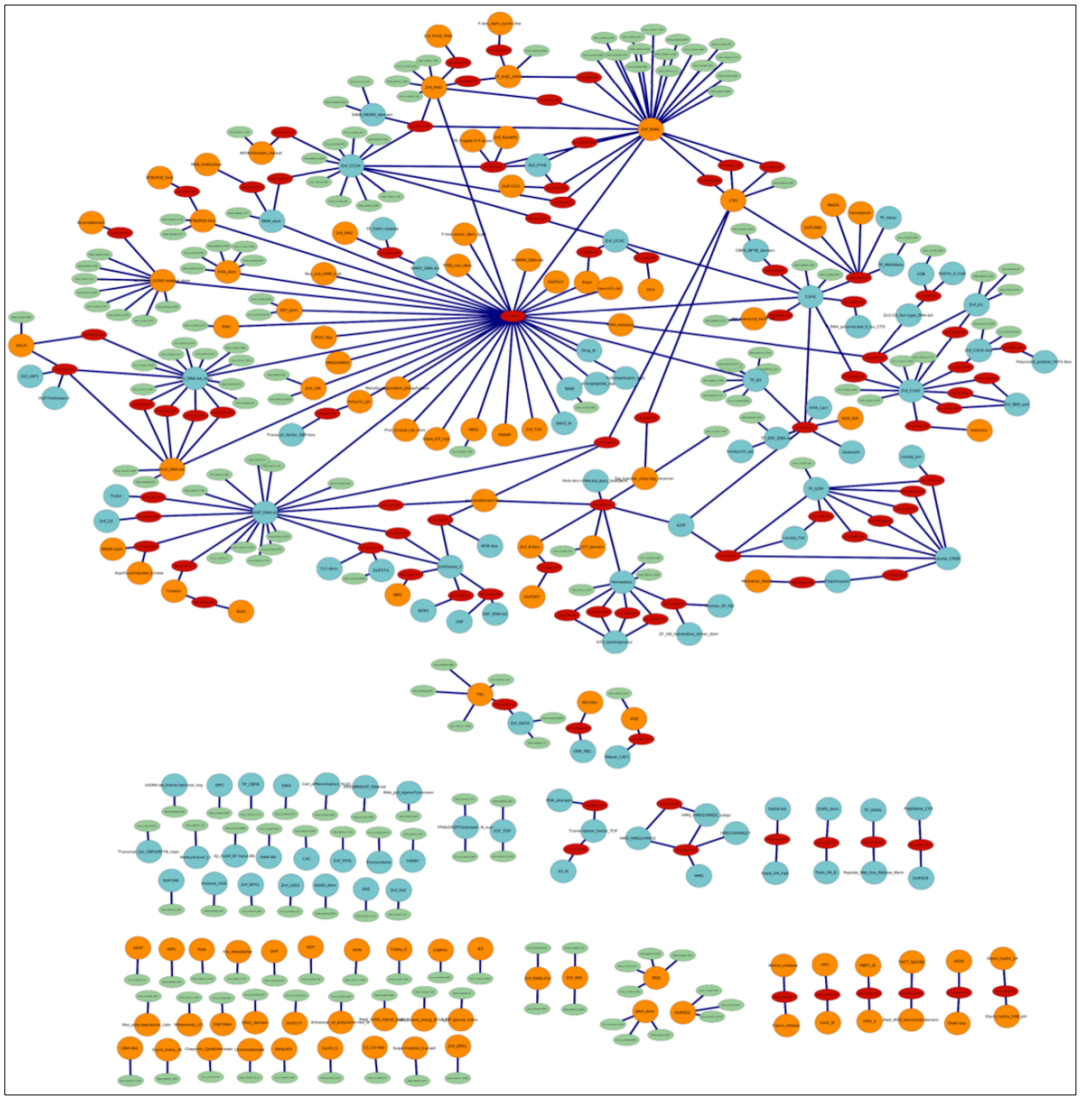


Figure 5. Distribution of TF and coregulators families among the different classes of more conservative networks.

Red nodes represent mixed and inclusive networks; the green ones depict exclusive and exclusive-distributed networks. Blue and orange nodes are TF and coregulator families respectively. Edges are connections among families and networks, as well as among the networks.

The classification of networks of paralogs containing more transcription factor and/or a coregulator families, was essential to understand the relationships among TFs and other genes, as genes sharing the same networks. This information is summarized in Figure 6 in which the structural relationships among transcription factor families (blue nodes), coregulator families (orange nodes), other type of genes and orphan TF genes, are shown. It's worth to notice the high connectivity among some families, and between TFs/coregulators and classes of other genes. Particularly, it is important the

sharing of the same network of TFs/coregulators belonging to annotated families and orphan genes. In fact, the identification of structural relationships among them, as well as the sharing of the same InterPro domain, can be useful to further refine their annotation.

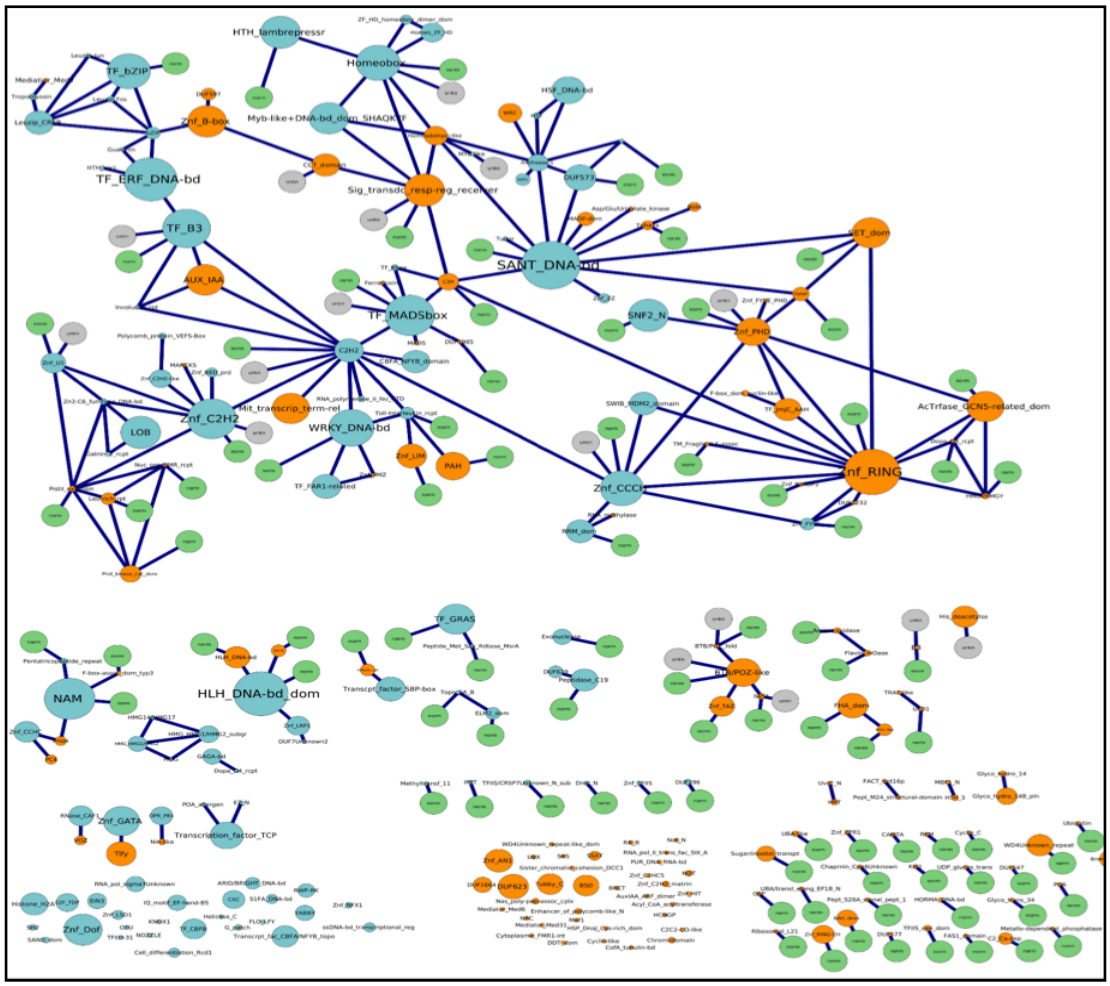


Figure 6. Structural relationships among TF and coregulator families and other classes of genes.

Transcription factor families are depicted by blue nodes while coregulator families by orange ones. Genes not annotated as TF/coregulators are indicated with green nodes, whereas grey ones depict orphan TF genes. The size of each node is correlated to the size of the family. The edges connect different families, based on their distribution among the more conservative networks.

4.3.2 Refining the annotation of Orphan transcription factors and coregulators

The 279 orphan genes are genes without a clear association to a family. Most of the

latter are coregulators (209 out of 279), while only the 25% (70 out of 279) are transcription factor genes. According to the classification listed in the previous paragraph, these genes are distributed in 20 exclusive networks, 5 exclusive-distributed, ones, 42 mixed and 80 inclusive networks. The lack of any structural relationship with transcription factor genes from other families of the orphan genes contained in exclusive networks makes further investigations aimed to refine their annotation limited to the fact that they represent a family just in case they are all crosslinked by paralogous relationships among themselves. By contrast, the sharing of a mixed network with one or more families of TFs and/or coregulators, represents the first step for the annotation of the orphan genes in case they share similarities with the vast majority of the members of one of the whole families. As explained in Chapter 2, networks are made of genes with at least one paralogy relationship. This means that not all the genes in a network are directly connected to each other. In this frame, we searched mixed and inclusive networks for the presence of orphan genes that share paralogy relationships only with TFs or coregulators belonging to the same families. We then looked for presence of common InterPro domains associated to the related genes: if the domain shared between an orphan gene and a TF or coregulator group of genes belonging to the same family, we assigned that gene to the same TF/coregulator paralog's family.

This analysis allowed the refinement of the annotation of 129 orphan genes, distributed in 14 families (Figure 7).

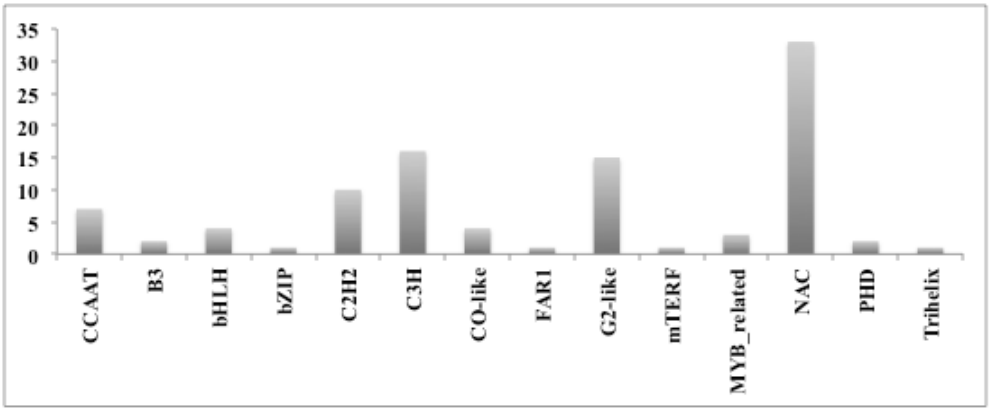


Figure 7. Orphan genes annotation.

The histogram shows the distribution of the re-annotated orphan genes (y axis) among 14 different gene families.

4.3.3 Single copy transcription factors and coregulators

As described before, transcription factors are considered dosage sensitive genes due to their key role in the transcriptional regulation. This implicates that they are retained after WGD events to maintain proper balance of proteins within transcriptional networks, but they tend to be eliminated following a smaller scale duplication (e.g. local and tandem duplicates, segmental duplicates, aneuploidy). At the light of these evidence, the presence of single copy transcription factor genes can represent an intriguing issue.

The analyses described in Chapter 3 allowed the identification of 3408 single copy *Arabidopsis thaliana* genes that were further investigated, in terms of ESTs validation and orthology relationships with other species. According to our results, we classified only 53 out of 2802 TFs/coregulators as single copy genes. In particular, 19 and 26 TFs and coregulator respectively, were classified as singleton genes confirmed by ESTs, whereas 8 out of 53 genes (3 transcription factors and 5 coregulators) were assigned to the not confirmed class.

45 singleton genes were orphan genes confirmed by ESTs, among the singleton genes with at least one ortholog in *O.sativa*, *S.bicolor*, *V.vinifera* and *P.trochocarpa* (listed in Supplementary material). We found that 42 out of 45 genes have at least one ortholog in all the considered species. Only 3 genes don't share orthologies with *O.sativa* and *V.vinifera*.

These results are in agreement with the gene balance hypothesis, since the presence of a high copy number of transcription factors in a genome which underwent whole genome duplication events

4.3.4 Transcription factor as part of the *Arabidopsis thaliana* paralogy network browser

The new TFs /coregulators classification was included in the implemented *Arabidopsis thaliana paralogy network database*, described in Chapter 2 and available on line at the address http://biosrv.cab.unina.it/athparalogs/main/index.

The query of a TF/coregulator gene, as well as of a TF/coregulator family, results in a table, which provides information about the genes (either in terms of paralogs or singletons) and their classification. The inclusion of these data in the database,

increments the usefulness of the resource provided, both for evolutionary studies and for the investigations on the organization of these genes.

4.4 Conclusion and discussion

Since *A. thaliana* is the reference genome for plant comparative genomics, a reliable annotation of its transcription factor genes is mandatory. The focus of this chapter was the identification of complete sets of *Arabidopsis* TFs and coregulators to fully understand their organization in a complex genome, but also to make a full use of *Arabidopsis* as a model plant for a comparative approach to annotation, TFs included. The second goal was the understanding of the attitude of TFs and coregulators in terms of duplicated and singleton genes in an ancient polyploidy organism.

4.4.1 The importance of a novel classification

Through the integration of reference databases first, and with the domain analysis then, we provide a solid classification of TFs and coregulators in *A.thaliana.* It's evident from our results that the integration of available databases is not enough to reach a comprehensive, reliable and unique classification. In fact, more than 30% of transcription factors were classified as putative TFs, since the heterogeneous information among the different databases. Moreover, a huge number of TFs were typical of each database. Our analysis allowed the removal of questionable features of putative genes and their re-classification as transcription factors or coregulators, Moreover, we found DNA-binding domains in proteins encoded by genes annotated as coregulators in other databases: this allowed the re-assignment of the latter in the TFs category. Our classification could not assign to a unique family 279 genes, annotated as "orphans". We overcame this issue considering the paralogy relationships in which these genes were involved in, exploiting the support of the network collection described in Chapter 2.

4.4.2 TFs and coregulators distribution in networks of paralogs

The second important goal achieved in this work is the analysis of the transcription factor's families in terms of paralogies: we identified structural relationships between TFs belonging to different families and/or between TFs and coregulators as well as relationships among TFs/coregulators and other classes of genes. As shown in Figure 5, TFs belonging to different families are included in the same network sharing one or more paralogy relationships with different classes of genes (TFs, coregulators, non-TFs). On the other hand we found exclusive families organized in exclusive networks, i.e networks involving only TFs belonging to the same family. Moreover, it may happen either that all the genes of the family are exclusively contained in one network (exclusive networks), or the genes are split among different exclusive networks (exclusive-distributed networks with exclusive families). Genes belonging to these networks don't share any structural relationships with other gene, indicating the specificity of the related gene family. On the other hand, a huge number of families are contained in inclusive networks, sharing paralogies with different classes of genes. Moreover, TFs and coregulators share paralogies also with orphan genes: the identification of such relationships together with the analysis of the InterPro domains associated to the related genes, allowed the refinement of the classification of 100 out of 279 orphan genes in 14 families, supporting the *Arabidopsis thaliana* annotation. Indeed, this result confirms the usefulness of the database we organized in the annotation of unknown or unclassified genes, as assessed in Chapter 2.

The analyses here presented also allowed the classification of single copy TFs, their presence still representing an intriguing topic in a highly duplicated genome. Single copy genes analysis revealed that the annotation of some single-copy TFs may be limited.

This analysis reveals still intriguing aspects of this genome annotation and sheds novel insights on TF families evolution, driving versus more reliable methodologies for large scale comparative genomics.

4.5 Material and methods

4.5.1 Publicly available TFs databases

There are many powerful genomic tools [4] to power the research in the field of plant genomics and proteomics. Databases sources for *Arabidopsis thaliana* are most extensive than for any other plant. In this study, we used and integrated data from four publicly available plant TF databases.

4.5.1.1 The Plant Transcription Factor Databases (PlantTFdb)

The Plant Transcription Factor Databases (http://planttfdb.cbi.pku.edu.cn/) [6] contains 53 319 putative TFs predicted from 49 species. We used the information about *Arabidopsis thaliana*, for which this dataset collected 2023 transcription factors classified into 58 families (TAIR9).

Strengths of this database are important information about each TF family going well in depth as far as giving the functional domains for each TF. Another positive aspect of this database is that it includes citations associated with each individual TF family. This database takes it a step further by displaying TF entries of what other databases like UniProt, RefSeq, and TransFac have. Information such as TF family and associated genes can be accessed by selecting a plant species and then browsing by family or by chromosome.

Once the TF family is selected the next choice is by gene. The gene information consists of some basic information, gene structure, annotation, protein sequence features, 3D structure, ontology, expression, sequence, and references to other databases. All the information given by selecting each gene is very well organized and visually appealing to the user. In the main homepage, it is important to note that each plant species has a database and a page of its own that opens when selected.

The BLAST option will provide different types of blast searches like blastp, blastn, blastx, and tblastn. For each BLAST option, it is possible to choose one of the 49 plant species. The search option takes to sub-databases by selecting specific species. Available downloads include protein sequences files, CDS (coding sequence) sequences files, and ortholog prediction files. The overall website is well organized, and maintained; last revised in October 22, 2010.

4.5.1.2 The Plant Transcription Factor Database (PlnTFdb)

The Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v2.0/) [7] currently contains 28193 protein models, 26184 distinct. PlnTFDB currently contains 2657 protein models, 2451 distinct protein sequences (TAIR8) of *Arabidopsis thaliana* used for our analysis.

The assortment of genes in each of the families is based on the presence of one or more characteristic domains previously described in the literature (identified through statistical analyses). To identify genes coding for transcription factors, previously constructed domain alignments (from the Pfam database version 23.0) or newly established alignments were used to query the Plant proteome, using the hmm pfam programs of the HMMER suite, links to the domain alignments are provided. Additionally, 279 proteins were categorized as orphans. These proteins contain one or more domain(s) whose presence, or combination, according to the literature, does not allow their classification into any of the defined families. Their role in the transcriptional regulation remains unclear.

Additionally, the database also provides information about genome databases, orthologs and co-orthologs, domain architecture (start codon, stop codon, and e-value), protein, and transcription sequences associated with each TF family.

4.5.1.3 AGRIS: AtTFDB - *Arabidopsis* transcription factor database

AGRIS (http://Arabidopsis.med.ohio-state.edu/AtTFDB/) (DAVULURI et al. 2003) was the first computational resource listing complete sets of TF in *A. thaliana*, Proteins were grouped into families according to their conserved DNA-binding domains in 1841 TFs classified in 50 families (TAIR9). Additionally, AGRIS lists putative cis-regulatory elements and links the TF information with putative target genes into gene regulatory networks. It used several different approaches to identify TFs and many families were identified through a domain search and a BLAST based approach. Publications were found through PubMed, and the conserved domain motif that characterizes each TF family here identified. Using the motif, a BLAST was conducted on the TAIR Web site, and the resultant sequences were then aligned and mismatches were discarded. Another approach, especially for large families where very few TFs had been identified, was an iterative BLAST approach. AGRIS was one of the motivations to develop a resource that encompassed a broad phylogenetic range

of plant species with sequenced genomes, like in other databases like, PlnTFDB.

4.5.1.4 The Database of Arabidopsis Transcription Factors (DATF)

The Database of Arabidopsis Transcription Factors (DATF) (http://datf.cbi.pku.edu.cn/) [11] is entirely dedicated to *Arabidopsis thaliana* TFs. It classify 1922 into 64 families (TAIR6) which are accessible from the homepage of the website. The family organization of all the TFs has multiple elements of the DNA-binding domain and neighbor-joining phylogenetic trees of each family. There is a link from the homepage on how to access other databases such as Plant Transcription Factor Databases, China (http://planttfdb.cbi.pku.edu.cn/). Information provided for a given TF family consists of the introduction, the structure, and the family binding sites. The following information can be downloaded: sequences of the family genomic sequence, CDS sequences, and protein sequences. Under the specific TF family, it is also possible to generate phylogenetic trees with genes in each TF family using the 'Phylogenetic Tree JPG Image' tool. One can also align the amino acid sequences encoded by genes of a TF family by choosing the 'Multi-alignment' option. A list of the total family 'loci' and 'gene models' are listed under each family. The gene models are organized in a table setup similar to the family class, making it easy to look through the information.

4.5.2 Integration of the available databases

We downloaded all the data about Arabidopsis thaliana transcription factor genes from the databases listed above. We then organized the data in a MySql based database in order to compare and integrate different information from different databases. We used as key field of the comparison the Locus id and we performed pairwise comparisons first considering each database in pairwise comparisons, and then altogether . After this integration, the genes were assigned to three different classes:

1. Confirmed TFs: those genes classified as transcription factors in all the considered databases. Since PlantTFdb is based on the last genome release and the pipeline implemented for the classification is consistent, we consider the latter as the more reliable one. Indeed,PlantTFdb and in at least another

database are considered as confirmed TFs.

2. Putative TFs: genes not annotated as transcription factors inPlantTFdb but present in at least on other database.
3. Co-regulators: Transcriptional regulator genes (annotated only in PlnTFdb)

4.5.3 InterProscan: DNA-binding and regulatory domain identification

The protein sequences (TAIR10) encoded from each gene classified by the integration of the databases, were downloaded from the TAIR website ().We used InterProScan package, version 34, specifically we used the following software: ScanRegExp, ProfileScan, FprintScan, HMMPfam, HMMPIR PIR, Superfamily. The obtained domains were further manually considered: we first checked for significant Gene Ontology terms (for example DNA-binding or protein binding or chromatin remodeling functions, ecc…). We also checked literature and the transcription factor encyclopedia []. The final classification of the selected InterPro domains, together with the gene annotations in the considered available databases allowed us to assign each gene to a particular TF or coregulator family. Only 279 genes were classified as orphans since not enough information was collected.

4.5.4 TF association to networks and singletons

Network and singleton results described in Chapters 2 and 3 were used to investigate on the duplication of TFs and coregulators. We first selected the networks, obtained with the two E-value thresholds $E \leq 10^{-5}$ and $E \leq 10^{-10}$, containing at least one of the TFs/coregulators present in our list. Only more conservative networks were used for further analyses ($E \leq 10^{-10}$) since our interest in the identification of stringent and reliable paralogies involving TFs. Cytoscape [] was used to visualize the relationships among TF/coregulator families and the distribution of the latter among networks.
Since some of the TFs/coregulators identified are not included in networks, we searched these genes among the singleton genes obtained with the pipeline described in Chapter 2. Information about EST validation and orthologies, were also included.

Chapter 5

The impact of alternative splicing on human transcription factor genes

5.1 Introduction

Cellular life must recognise and respond appropriately to diverse internal and external stimuli. By ensuring correct expression of specific genes at the appropriate times, the transcriptional regulatory system plays a central role in controlling many biological processes: these range from cell cycle progression and maintenance of intracellular metabolic and physiological balance, to cellular differentiation and developmental time-courses [1-2]. As described previously, transcription factors (TFs) represent key components of the regulatory machinery, binding the DNA and interacting with specific genomic regions in the neighbourhood of the target gene, and so promoting or supressing its expression [3]. Numerous diseases result from a breakdown in the regulatory system, and a third of human developmental disorders have been attributed to dysfunctional TFs [4-5].

The regulatory activities of TFs are themselves controlled at two levels: the transcriptional, which is determined by other TFs, and post-transcriptional steps including alternative splicing [8], i.e. the process by which the exons of a primary gene transcript are reconnected in multiple ways [10]. In humans, ~95% of multi-exonic genes are alternatively spliced [11]: thus alternative splicing provides a potentially widespread and important mechanism for controlling the regulatory activities of TFs (Figure 2) in different tissues [12]. In this frame, the comprehension

of the regulation of human transcription factor genes in terms of alternative splicing may represent a starting point not only for a better understanding of the TF regulatory machinery, but also for the investigation of human diseases due to a defective TF activity.

In fact, most studies equate the transcription of TFs with regulatory activity; however there are important examples of TFs whose activities are known to be modulated primarily at the post-transcriptional level. An example is the *era*-1 gene: two alternatively spliced mRNAs are produced, one of which encodes the active form, while the other produces a protein lacking the DNA-binding domain (DBD). This truncated isoform is incapable to binding to DNA and activating gene expression [9].

So far, no study has examined the extent and impact of alternative splicing on transcriptional regulation in any mammalian genome. Without this information, we are unable to understand the gene-expression programmes that allow cells to take on their individual identities. In this introduction, some examples of how alternative splicing can affect transcription factor activity will be provided in order to better understand the aim of the analyses described in this Chapter.

5.1.1. TFs and alternative splicing: some examples.

Transcription factor genes are highly modular in the organization of sequence elements required for DNA binding, dimerization, ligand binding, subcellular localization and transcriptional activation. A wide range of alternative splicing strategies operates on this modularity to generate a variety of protein isoforms from a single gene. Alternative splicing can affect TF structure in different ways [15]: alterations can be in the DNA-binding domains affecting their affinity or specificity; or alterations can modulate interactions of TFs with their cofactors (activators and repressors).

In particular, alternative splicing can generate different isoforms of a particular transcription factor, in different ways:

- **Production of active and inactive isoforms**: more alternatively spliced mRNAs are produced, some of which encodes the active form, while the others produce a protein lacking the DNA-binding domain (DBD). The truncated isoforms are incapable to binding the DNA and activate gene expression [1][2] (Figure 1A).

- **Production of active and inactive isoforms with different effects on transcription:** among the alternative isforms, some encode the active form, whit either the DBD or the co-regulator domain, while the others produce a protein lacking of the co-regulator domain. This truncated isoform is capable to binding to DNA but unable to activate gene expression. The isoforms compete for the same binding domain [4][5] (Figure 1B).

- **Production of active isoforms with different binding domains in different cell types:** the alternative mRNAs are produced with different specificities in different cell types [5]. An example is the Oct2 gene: this gene is present in B lymphocytes but it is absent in most other cell types, playing a critical role in the B-cell-specific transcription of the immunoglobulin genes. Whereas, the B-cell protein activates octamer-containing promoters, the neuronal protein inhibits octamer-mediated gene expression. [6] (Figure 1C).

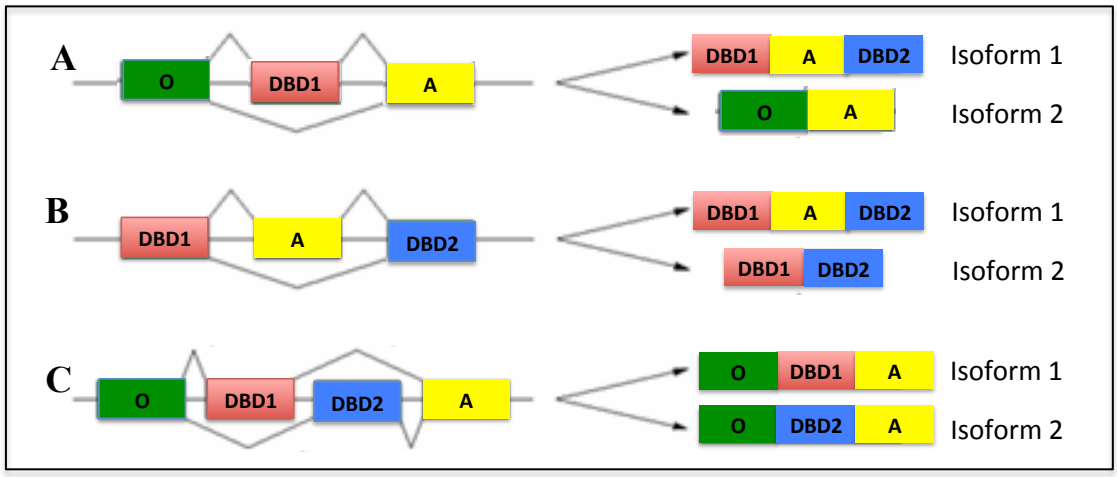


Figure 1. Alternative splicing mechanisms act on TFs modularity.

A. Two different isoforms are produced by AS mechanism: isoform 1 is the active one with both DBDs and the activation domain. On the buttom the inactive isoform (isoform 2): it can bind to DNA but cannot activate transcription since it lack of the activation domain. Therefore this isoform inhibit transcription by competing for binding to DNA with the activating isoform. **B.** Isofomrs 1 and 2 are both active but have different specificity for different target genes.

Active isforms with different properties (different specificities) can also be produced by the use of different splice site. The isoforms are active but with different specificities [3].

All these studies have case by case basis [1-14], and relatively little is known about the global regulatory properties of AS on TF genes.

5.1.2 Aim of the chapter.

In this chapter, we performed a genome-wide investigation on alternative splicing of TFs in the human genome. In particular, I will determine the impact of alternative isoforms on a TF's regulatory activity, assessing the extent to which different TFs are alternatively spliced. In other words, the main aim of this work is the achievement of a more complete insight into the complexity of the regulation of TF activity by alternative splicing mechanisms.
In the previous chapter we analysed TFs in term of duplicated genes. Gene duplication and alternative splicing are distinct evolutionary mechanisms but both provide the raw material for new biological functions and for gene amplification. In this chapter we present also preliminary results about the relationships of these two biological mechanisms in *Homo sapiens* and *Arabidopsis thaliana*.

5.2 Results

5.2.1 Transcripts integration and identification of TFs alternative splicing

The main focus of this work is the analysis of the impact of alternative splicing events on human TFs. To fulfill these aim, we performed a genome wide investigation of the human transcription factor genes, identifying all the alternative splicing mechanisms and dividing them into two classes: one isoform and more isoforms TFs. The main steps of our analysis are summarized in the pipeline in Figure 2.

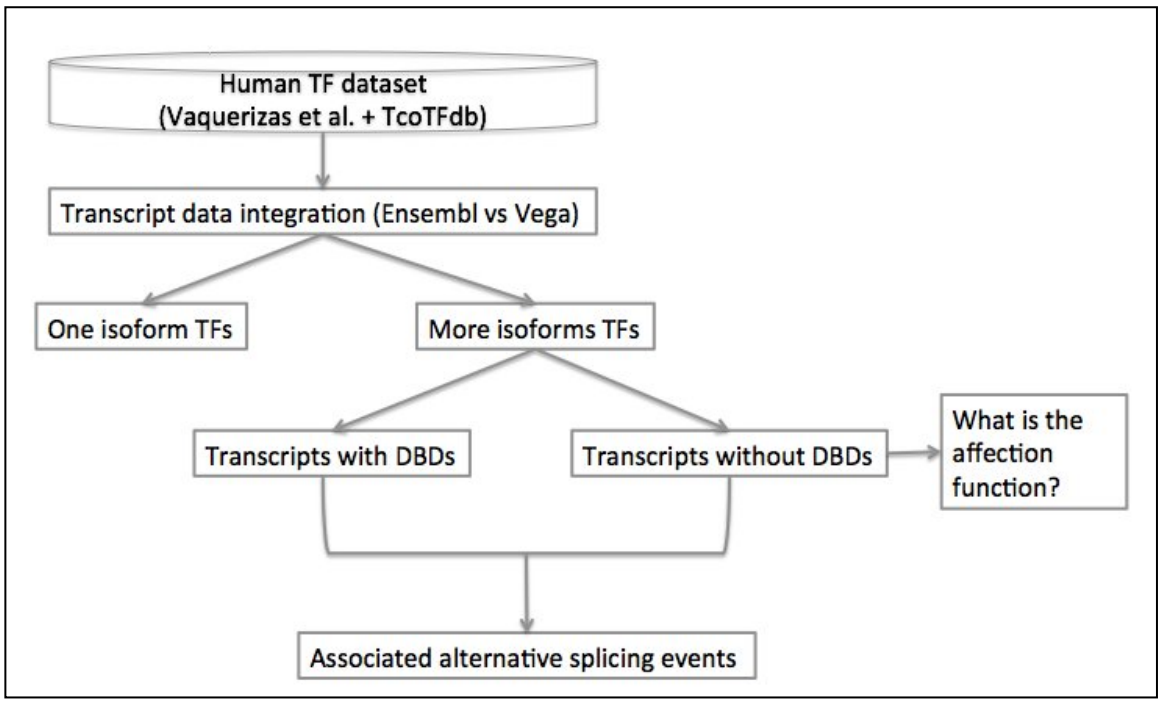


Figure 2. Pipeline: the main step of the analysis

This pipeline summarizes the main steps of our analysis. Starting from the list of human TFs, we integrated the transcript data in order to divide TFs in one isoform and more isoform genes. Transcripts encoded by the more isoform TFs were further divided in transcripts with DBDs and without DBDs. The related alternative splicing events were identified in order to understand the impact of alternative splicing on the presence of the DBD.

5.2.1.1 Human transcription factors dataset.

The first step of this analysis consists in the identification of TFs that are associated with more than one transcript, suggesting potential splicing events. To this aim, the correct identification of transcription factor genes and their transcript sequences is mandatory. For the analysis of Arabidipsis thaliana TFs (Chapter 4), we made the effort to produce a new classification of this class of genes, since the lack of a reference dataset.

To study human TFs instead, we rely a landmark study by the Luscombe Group [7], which reported that the human genome encodes for 1,391 TFs organized in 23 families. In addition to this, we exploit TcoF-D [Schaefer et al., 2010], Dragon Database for Human Transcription Co-Factors and Transcription Factor Interacting Proteins (TcoF-DB). The latter is totally based on the Vaquerizas et al. data [] but extended the dataset to 1418 TFs, extracting 27 proteins from TRANSFAC [Matys V,

et al., 2003] and TFCAT [Fulton DL, et al., 2009]. Since 31 out of 1418 were retired in the new genome release, the final dataset consists of 1387 human TFs. In total, we used 1387 out of 1418 human TFs, since 31 genes were retired in the new genome release.

5.2.1.1 Transcript data integration and analysis

The second step of the pipeline consists of the transcript data integration, i.e. the identification of the mRNA sequences encoded by human TFs. To this aim, we used both the Ensembl [] and the Vega databases []. The first one describes approximately 178,191 transcripts that have been identified either through automatic annotation, i.e. genome-wide determination of transcripts, or manual curation, i.e. reviewed determination of transcripts on a case-by-case basis. Instead, the Vega database describes 152,290 transcripts annotated only manually. Among the list of 1387 human TFs, we found transcripts information for all of them using Ensembl database, while information about 102 TFs are missed using data annotated in the Vega database.
The integration of transcript data from the two databases was outstanding, since to good overlap between the obtained results (in terms of number of transcripts per TF). Since Vega database lacks of the annotation of 102 TFs and Ensembl database is considered a standard database for genomic features annotation, we eventually decided to use Ensembl data.
Ensembl classifies gene transcripts according to their Ensembl Biotype, as follows:

- protein coding
- processed transcripts (non protein coding, Long intergenic non coding RNA, retained intron)
- pseudogene
- Non sense mediated decay (NMD)
- Artifact.

Only 1349 out of 1387 TFs have at least one protein-coding transcript, while the remaining ones have only non protein coding transcripts: in particular we found 2 genes coding for Long intergenic ncRNA, 22 pseudogenes and 14 genes coding for other processed transcripts. These genes were discarded from our dataset, reducing the number of protein-coding TFs to 1349. Figure 3 shows the numbers of non protein-coding transcripts encoded by the latter: it is worth to note that the great part

of non protein-coding TF transcripts are classified as retained intron or belong to non sense mediated decay (NMD) class (Figure 3). NMD was recently postulated as one of the most important quality-control mechanisms with many functional implications concerning the stability and biological relevance of splice variants.

In fact, alternative splicing mechanisms may often regulate gene expression by activating the NMD mechanism. In particular, approximately 5% of all human alternative splicing events produce variants that belong to the NMD class, containing premature termination codons (PTCs), which can be recognized, targeted and efficiently degraded by NMD (if the coding sequence of a transcript terminates > 50bp from a downstream splice site then it is tagged as NMD).

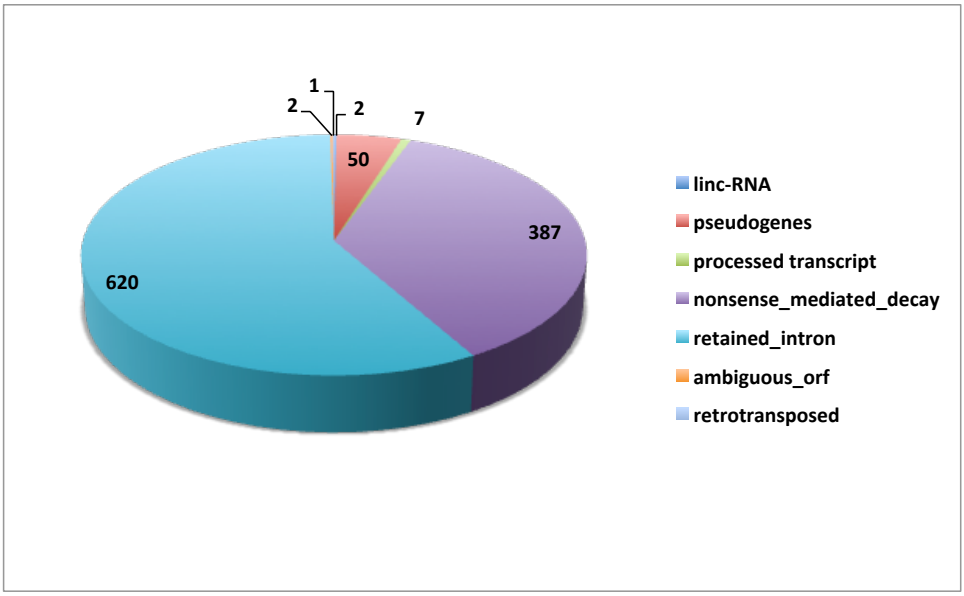


Figure 3. Non protein-coding transcripts identification.

The diagram shows the number of TF non protein-coding transcripts for each Ensembl biotype.

5.2.1.2 TFs classification: one isoform and more isoforms genes.

In order to classify the 1349 TFs on the bases of the number of their isoforms, we focused the analysis only the protein-coding transcripts encoded by each TF. We found that the 28% of the TFs encode for only one protein-coding isoform, while the 72% of them are classified as more isoforms TFs.

We repeated this analysis also on all the other human genes, to assess whether TFs are

subject to greater numbers of splicing events in comparison with non-TF genes. We found that the percentages of one isoform and more isoform human genes is almost the same considering also the genes not annotated as TFs, indicating that TFs are not subject to more alternative splicing events in comparison with other genes (Table 1).

Transcription factors		**All genes**	
One Isoform	More Isoforms	One Isoform	More Isoforms
379 (28%)	970 (72%)	4257 (20%)	16908 (80%)

Table 2 Transcription factors classification: One isoform and more isoforms human genes

The table depicts the number of TFs one isoform and more isoform genes. In bracket the percentage of each class calculated on the total number of TFs. The right, the same results are shown considering all the human genes.

We compared transcript sequences with the correspondent gene structure, to investigate on the number of exons in order to see which exons have been included or excluded (cassette exons). In doing so, TFs were further divided as single-exon TFs and multiple-exons TFs. The results are summarized in Table 2. As expected, all TFs classified as more isoforms ones present more than one exon. Indeed, 56 out of 379 one isoform genes present only one exon, whereas the remaining 323 one isoform TFs were classified as multiple exons genes. We further analyzed the one isoform TF's exons and we found that all of them are constitutive exons (i.e. exons that are always included in the mature mRNA). This evidence confirms the non-spliced nature of these genes.

One Isoform		More isoforms	
Single exon	Multiple exons	Single exon	Multiple exons
56	323	0	970

Table 2. Single and multiple exons TFs.

A second classification of one isoform and more isoforms TFs on the bases of the exon number is depicted in this table. It's worth to note that 323 out of 323 one isoform multiple exons genes are composed by constitutive exons.

Once TFs were classified as one isoform or more isoforms genes, we assessed whether different TFs families have a different number of genes belonging to these two categories. To this aim we used the classification from Vaquerizas et al. [] in which TFs (TFs) were arranged into families according to their DBD composition. In particular, they defined 23 transcription factor families, based on the presence of 347 different DBD domains. According to Vaqerizas et al [], 594 out of 1387 TFs belong to the ZNF-C2H2 family. We decided to further classify genes belonging to this family, on the basis of the presence of the regulatory domain. In particular, ZNF-C2H2 genes were sub-classified for the presence of the typical conserved regulatory regions, such: KRAB, SCAN and BTB/POZ. Table 3 indicates the final list of gene families together with the total number of genes herein contained and the classification of the latter in one isoform and more isoforms genes.

Family	Genes	One isoform	More isoforms
AP2	5	2	3
BZIP	49	15	37
CP2	6	0	6
DM	7	2	5
ETS	27	3	24
Forkhead	43	22	21

Heat-shock-Other	6	0	6
HLH	89	34	55
HMG	32	10	22
Homeodomain	233	99	134
IPT/TIG	17	0	17
IRF	9	1	8
MAD	10	0	10
MADs-box	4	0	4
NHR	47	2	45
Other	65	12	53
P53	30	3	27
POU-Homeodomain-Other	15	6	9
RFX-Other	7	1	6
SAND	10	1	9
TDP-Other	9	4	5
ZNF-C2H2 $(C2H2)_X$	236	55	181
ZNF-C2H2 BTB/POZ	47	11	36
ZNF-C2H2 KRAB	264	45	219
ZNF-C2H2 SCAN	21	6	15
ZNF-C2H2 SCAN-KRAB	26	0	26
ZNF-GATA	9	1	8

Table 3. Transcription factor families final classification.

The table shows the number of genes contained in each transcription factor family and their classification in one isoform and more isoforms.

It is worth to note that the amount of more isoform genes is different in each family, suggesting that the number of isoform is not indicative of a certain class of TFs.

5.2.1.3 Identification of alternative splicing events affecting the presence of DNA-binding domains.

We found at least one alternative splicing mechanism for 1017 out of 1349 TFs. In Table 4, the total numbers of the identified events are indicated.

Splicing event	Total events	Only protein coding transcripts
alternative 3' splice site (A3SS)	559	423
alternative 5' splice site (A5SS)	542	421
alternative first exon (AFE)	1640	1303
alternative initiation (AI)	1033	798
alternative last exon (ALE)	376	296
alternative termination (AT)	1153	897
cassette exon (CE)	3038	2377
constitutive exon (CNE)	2584	2513
exon isoform (EI)	12	9
intron isoform (II)	1299	996
intron retention (IR)	1633	1143
mutual exclusion (MXE)	330	278

Table 4. Alternative splicing events indentification.

The table depicts the type of alternative splicing events affecting TFs. The second column indicates the number of total events detected considering all the TF transcripts. The third

column depicts only the AS mechanisms identified for protein-coding transcripts.

Since our goal is the extent to which alternative splicing regulates transcription factor activity, we decided to focus our analyses only on alternative splicing mechanisms involved in the lost of the DBD. In fact, as described before, TFs perform their regulatory activity binding the DNA interacting with specific genomic regions in the neighbourhood of the target gene. The lost of such domains in some of the isoforms encoded by TFs represents an interesting issue for the understanding of the transcriptional regulatory activity.

To this aim, we detected, for each more isoform transcription factor protein, the DBD position, its genomic coordinates on the gene and the exon/s coding for it. Then we checked for the presence/absence of the domain in each TF transcript, taking into account the correspondent alternative splicing event.

This analysis allowed the classification of TF transcripts in two categories: i) transcripts with at least one DBD; ii) transcripts without DBDs. The numbers of transcripts belonging to the two classes of transcripts for each TF family, are indicated in Table 5.

Family	**All transcripts**	**Without DBD**	**With DBD**
AP2	21	6	15
bZIP	208	52	156
CP2	27	8	19
DM	16	2	14
ETS	130	32	98
Forkhead	123	25	98
Heat-shock	31	3	28
HLH	324	63	261
HMG	180	48	132
Homeodomain	764	159	605
IPT/TIG	112	9	103
IRF	53	4	49

MAD	62	5	57
MAD's box	31	1	30
NHR	309	52	257
Other	306	63	243
P53	150	19	131
POU	47	13	34
RFX	49	18	31
SAND	46	11	35
TDP	48	33	15
ZNF-C2H2 (C2H2)x	1017	129	888
ZNF-C2H2 BTB/POZ	175	26	149
ZNF-C2H2 KRAB	982	24	958
ZNF-C2H2 SCAN	58	7	51
ZNF-C2H2 SCAN-KRAB	101	32	69
		15	33
ZNF-GATA	41	14	27

Table 5. Lost of DBDs in different families.

For each TF family, the numbers of transcripts with and without DBDs together with the total number of transcripts are shown.

To better investigate on the function of the 1042 transcripts for which we didn't detect DBDs, we searched the correspondent proteins for the presence of other types of InterPro domains. These domains were manually investigated, examining the description and the associated literature. The 26% of these proteins have at least one protein-binding domain, suggesting that most probable they are inactive isoforms: they can still interact with the coregulators but without interact with DNA (Figure 1A). Interestingly, the 41% of the proteins without the DNA-binding domain, present domains RNA-binding activity, suggesting a double function of the correspondent gene in terms of binding DNA and RNA using different protein products. For the

remaining 33% of the proteins no significant functions were detected, since the presence of general domains, classifies as “Other”.

5.2.1.4 Different DNA-binding domains within the same family.

The 4470 transcripts coding for at least one DBD were further investigated, in order to assess whether DBDs in proteins belonging to the same family are identical or share some sequence differences. To this aim, we perform a multiple alignment of the protein sequences of DBDs belonging to the same TF family.

We found that only some families have proteins with identical DBDs, suggesting that these isoforms are different in terms of coregulatory domains. The great part of the families is composed by isoforms with different DBDs in terms of aminoacid sequences. For each family, a phylogenetic tree for isoforms with differences in DBDs was inferred. This analysis allowed the classification of sub-families within each transcription factor family, based on sequence similarities among the DBDs. This evidence was confirmed through a motif analysis, performed using MEME suit: isoforms belonging to the same sub-family share same motifs, while transcripts found in different sub-families (within the same family) are depicted by different motifs.

An example of this results is shown for the AP2 family (Figure 4): it is composed by 2 one isoform and 3 more isoform genes. Among the 21 protein-coding transcripts belonging to this family, 17 encode for the AP2 DBD domain. The phylogenetic tree shown in Figure 5 depicts the presence of 3 AP2 sub-families. The presence of 3 different subfamilies is also confirmed by the presence of different motifs among the different AP2 sub-families.

Figure 4. Identification of sub-families based on DBD sequence similarities.

The example shows the phylogenetic tree obtained by the multiple alignment of the AP2 DBD protein sequences, within the AP2 gene family. The tree shows the presence of three sub-groups (depicted by square of different colours) of proteins within the same family. The three sub-groups were confirmed when the motif search analysis was conducted (on the bottom), using the AP2 DBD protein sequences. Ten different motifs were found, indicated by different colours (see legend on the top). The sharing of different motifs allowed the detection of the same sub-groups of proteins.

Family	**More isoforms genes**	**Genes with identical DBD**	**Genes with different DBD**
AP2	5	2	1
BZIP	49	5	32
CP2	6	4	2

DM	7	3	2
ETS	27	24	0
Forkhead	43	15	6
Heat-shock-Other	6	3	3
HLH	89	40	15
HMG	32	15	7
Homeodomain	233	55	79
IPT/TIG	17	15	2
IRF	9	7	1
MAD	10	8	2
MADs-box	4	3	1
NHR	47	32	13
Other	65	35	18
P53	30	15	12
POU-Homeodomain-Other	15	8	1
RFX-Other	7	3	3
SAND	10	6	3
TDP-Other	9	3	2
ZNF_C2H2_(C2H2)X	236	98	83
ZNF_C2H2_BTB/POZ	47	16	20
ZNF_C2H2_KRAB	264	120	99
ZNF_C2H2_SCAN	21	6	9
ZNF_C2H2_SCAN-KRAB	26	12	14
ZNF_GATA	9	7	1

Table 6. DNA-binding domain differences

More isoforms genes classification on the bases of protein sequence similarities among the DBDs. The number of genes with identical DBDs or with different DBDs are indicated for each family.

5.3 Alternative splicing and gene duplication: preliminary results of a comparative analysis between two reference species.

Alternative splicing and gene duplication are two important mechanisms for the increment of the gene amplification and in turn genome complexity.

We decided to perform a comparative analysis between *A.thaliana* and *Homo sapiens* in terms of number of isoforms and number of paralogs.

Since in the previous analyses we studied A.thaliana genome in terms of duplicated genes and *Homo sapiens* in terms of alternative splicing, we performed the complementary analyses in order to have the same information in both species.

In particular, in *A.thaliana* we counted the number of isoforms for each protein-coding gene first and for only the transcription factors then. This allowed the classification of the genes in one isoform and more isoforms ones (Table 7).

It's worth to note that the percentage of genes belonging to the two different classes it is completely different between *A.thaliana* (Table 7) and *Homo sapiens* (Table 1): according to our results the number of human more isoform genes is much higher than the one isoform ones. In *A.thaliana* we found an opposite behaviour. The same evidence arise for transcription factor genes. For this analysis we considered only transcription factor genes without coregulators.

Transcription factors		**All genes**	
One Isoform	More Isoforms	One Isoform	More Isoforms

1195 (75%)	387 (25%)	21607 (79%)	5809 (21%)

Table 7 Transcription factors classification: One isoform and more isoforms Arabidopsis genes

The table depicts the number of TFs one isoform and more isoform genes. In bracket the percentage of each class calculated on the total number of TFs. The right, the same results are shown considering all the *Arabidopsis* genes.

On the other hand, we found human paralogy relationships using Ensembl Biomart []. We found 14620 out of 21165 singleton human genes while 6545 out of 21165 resulted duplicated. Also in this case, an opposite behaviour between the two species was detected in terms of duplicated genes.

The achievement of these data allowed us to compare the two mechanisms counting the number of one isoforms and more isoform genes among the singletons and duplicated genes respectively, either in *Arabidopsis* or in human. The results are shown in Figure 5. The number of human duplicated genes among the one isoform ones is higher compared to the number of singletons. Viceversa, in *Arabidopsis* we found a higher number of one isoform genes among the singleton ones. These results suggest a different usage of the two mechanisms in the two species.

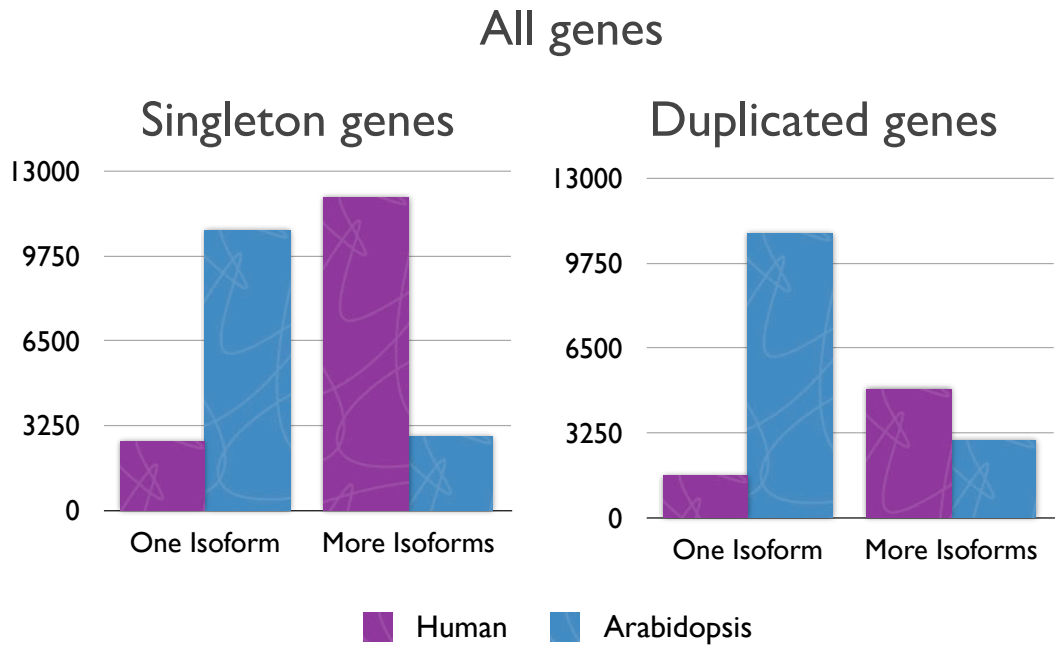


Figure 5. Alternative splicing versus gene duplication considering all *Arabidopsis*

and human protein-coding genes.

On the left, the histogram shows the number of one isoform and more isoform genes among the human (pink) and Arabidopsis (blue) singletons. The same information is shown in the histogram on the right, with the number of one isoform and more isoform genes among the human (pink) and Arabidopsis (blue) duplicated genes.

The same results were obtained considering only the transcription factor genes of the two species (Figure 6). In this case the number of human and Arabidopsis more isoforms genes are more similar either among singletons or the duplicated genes.

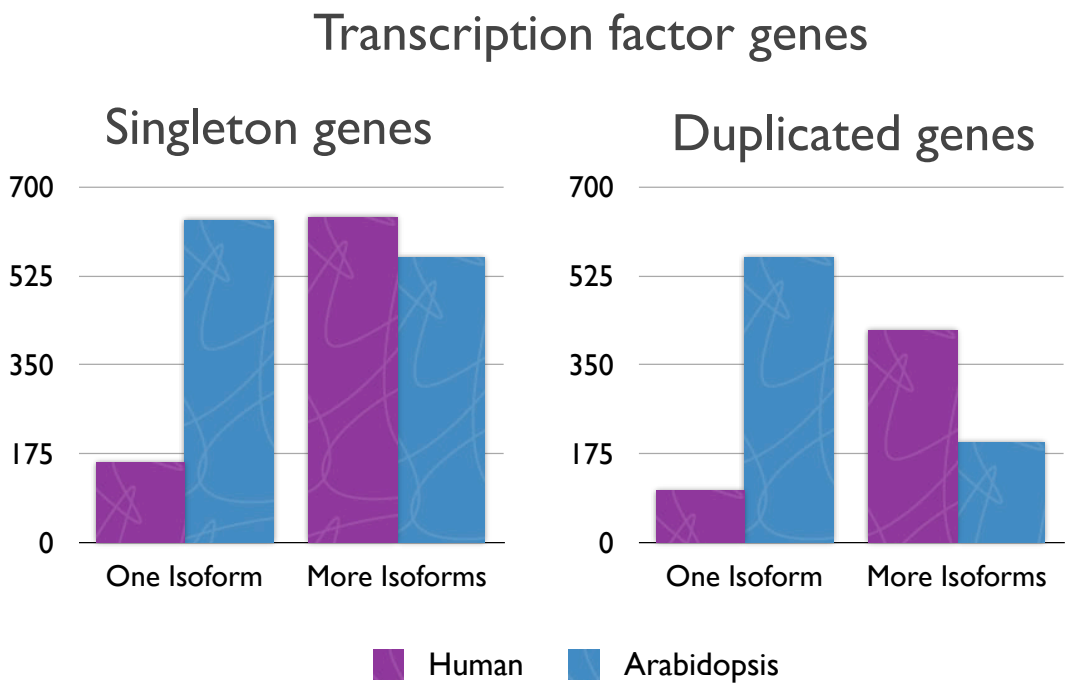
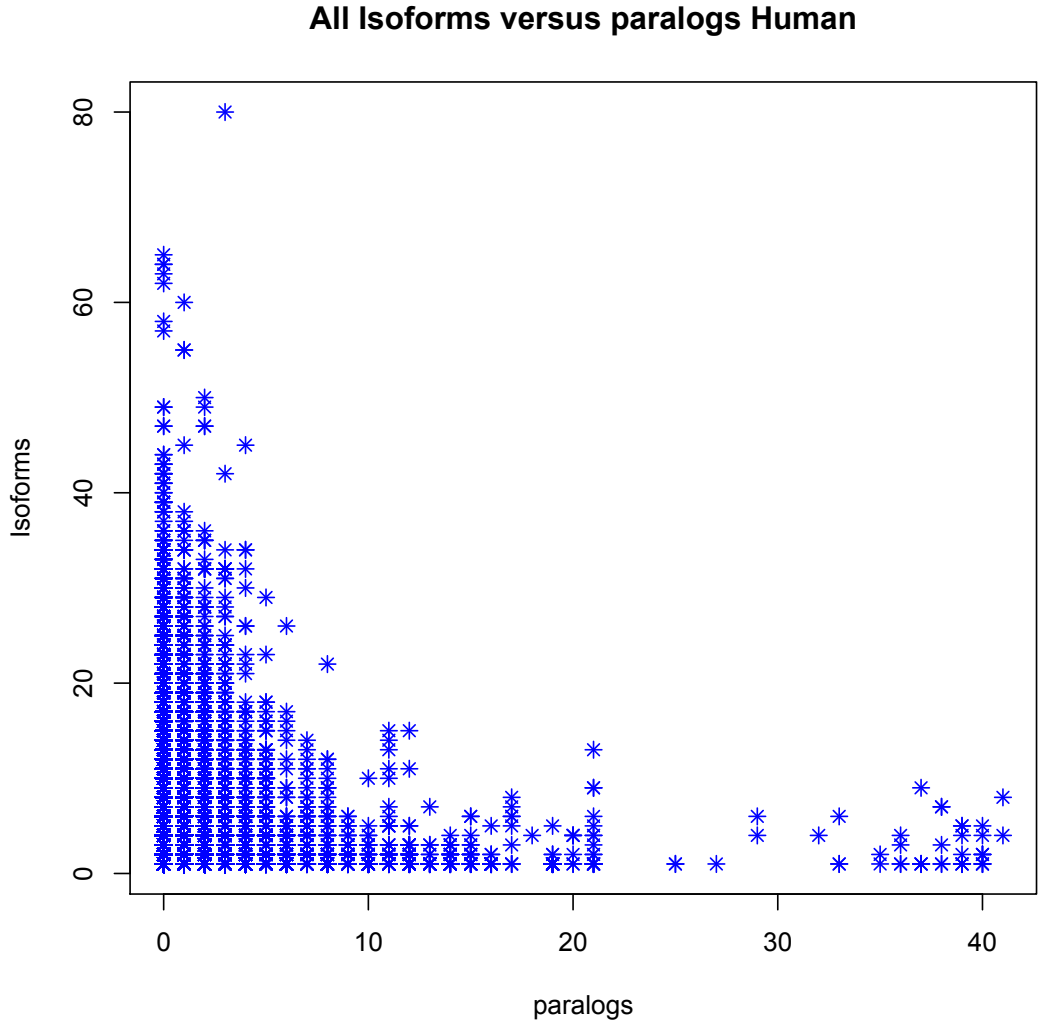


Figure 6. Alternative splicing versus gene duplication considering all Arabidopsis and human protein-coding genes.

On the left, the histogram shows the number of one isoform and more isoform genes among the human (pink) and Arabidopsis (blue) singletons. The same information is shown in the histogram on the right, with the number of one isoform and more isoform genes among the human (pink) and Arabidopsis (blue) duplicated genes.

Finally, the two species show a completely different correlation between number of paralogs and number of isoforms, as shown in Figure 7. At a first glance seems that the two species have use the two mechanisms in an opposite way to reach genome complexity. These are only preliminary results that reserve a deeper investigation.


All Isoforms versus paralogs Human
Isoforms
paralogs
0
20
40
60
80
0
10
20
30
40

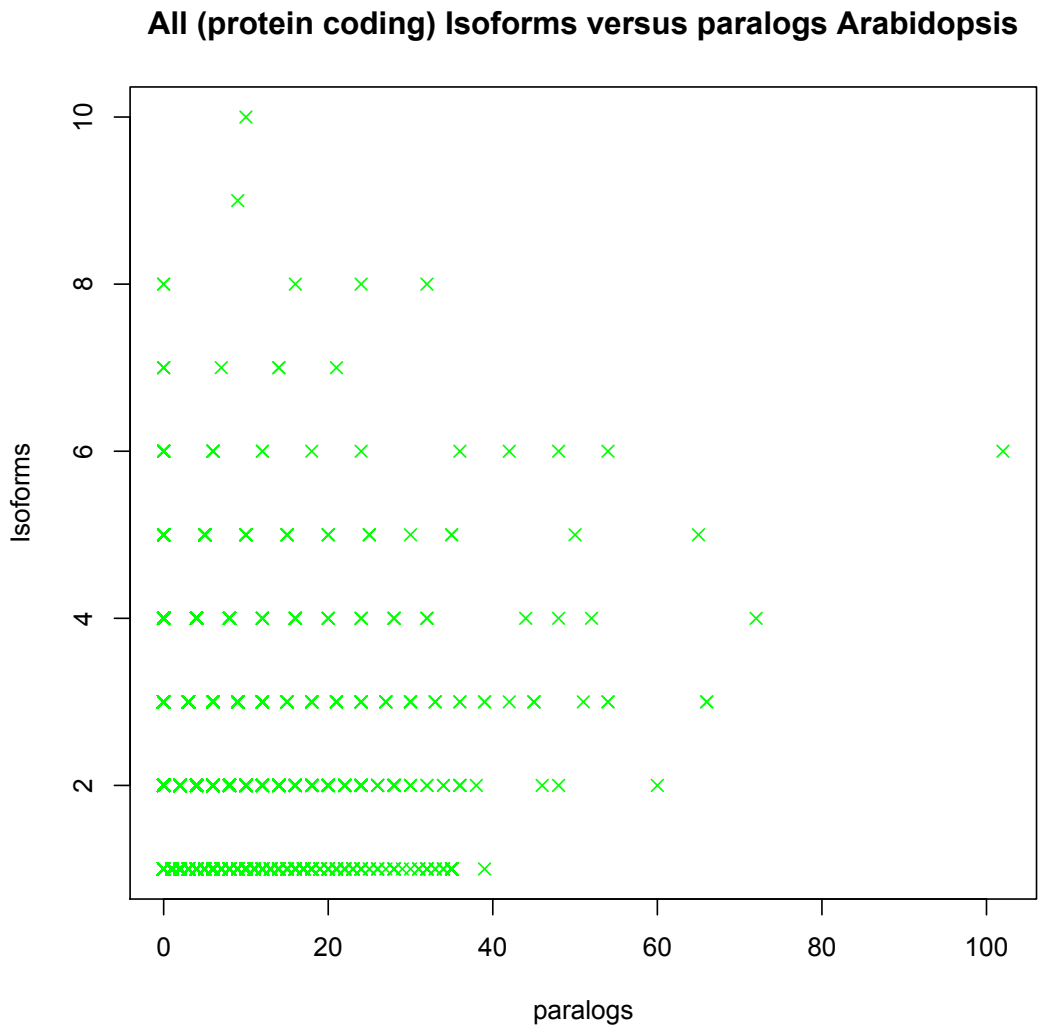


Figure 7. Number of isoforms versus number of paralogs.

This plots show the correlation between the number of isoforms (y axis) and the paralog numbers. The plot on the top shows the results relative to human (blue stars) while on the bottom the results relative on *A.thaliana* are shown (green stars).

5.4 Conclusion and discussion.

Alternative splicing of transcription factor genes is known to have a big impact on their regulatory function, particularly in higher organisms such as humans. However, apart from individual examples studied using molecular techniques, the extent and complexity of transcription factor splicing is still poorly understood on a genomic scale. Any understanding of the dynamics of gene regulation – particularly in specific

tissues and during time-courses – requires a comprehensive analysis of this phenomenon.

The presented analysis provides the first genome-scale study in any mammalian organism of how TFs are alternatively spliced and its impact on the transcription factor activity. In particular, our analyses are focused on the alternative splicing mechanisms acting on the modularity of transcription factor genes.

The presence of at least one DNA-binding domain in transcription factor proteins is request for the normal TF activity. We found a high number of TF isoforms without any DNA-binding domain, identifying the correspondent alternative splicing mechanism, which bring to the lost of the DBDs. All the results we showed are described considering the TF family classification.

Moreover, our analysis allowed the identification of transcription factor subfamilies, within each family. This is an important result that gives a surplus value to the family classification based on the presence of specific DBDs.

Furthermore we refined the TF family classification, considering the presence of different coregulatory domains among the ZNF-C2H2 proteins.

In addition, our results represent a starting point for the understanding of the dynamic usage of splice transcription factor isoforms and their regulatory impact in different cell types and along time-courses of important biological processes like differentiation.

Recent developments in transcriptomics (ie, the use of high-throughput sequencing to detect transcripts), allows us to detect different splice isoforms in different cell types. Our results can be used as template to use together with expression data to examine which types of TFs are expressed in given tissue types.

Many human diseases are linked to a breakdown in the normal regulatory function of TFs. For instance, targeting of the wrong genes for expression might result in the uncontrolled growth and cell division observed in tumour development [15]. In this frame the use of our results followed by the analysis of that are uniquely expressed in cancers, which are not normally observed.

5.4.1 Comparison between alternative splicing and gene duplication: interesting preliminary results.

The preliminary results we obtained comparing alternative splicing and gene

duplication events in *A.thaliana* and *Homo sapiens* are really interesting and pose intriguing question that can be further investigated.

First of all we detected a completely different number of genes with only one isoform and genes coding for different isoforms. In the same time, it is also different the number of duplicated genes found in the two species: A.thaliana tends to use gene duplication to achieve proteomic diversification since it contains a high number of duplicated genes and the great part of its genes have only one isoform or at most few isoforms. By the contrary, human genes have a huge number of protein-coding isoforms but the most part of them are in single copy.

These evidence must be validated and deeply investigated through more solid analyses but can be considered as a starting point to the understanding of the relationship between alternative splicing and gene duplication.

5.5 Material and methods

5.4.1 Human transcription factor dataset

Transcription factor genes were obtained from Vaquerizas et al. [Vaquerizas et al., 2009] and from TcoF-D [Schaefer et al., 2010]. The first dataset consists of 1391 TFs obtained looking for proteins that bind DNA in a sequence-specific manner. The second one relied on data previously published by Vaquerizas et al. and identified other 19 TFs proteins from TRANSFAC [Matys V, et al., 2003] and 8 from TFCAT [Fulton DL, et al., 2009]. The final list consists of 1418 human transcription factor genes. Genomic and protein information were obtained both from Ensembl and Vega databases. The first one describes approximately 178,191 transcripts that have been identified either through automatic annotation, i.e. genome-wide determination of transcripts, or manual curation, i.e. reviewed determination of transcripts on a case-by-case basis. Instead, the Vega database describes 152,290 transcripts annotated only manually.

Also all the information about TFs, such as coordinates, exon structures, exon number, were obtained using Ensembl BioMart [] (Human genome release GRCh37.p5).

5.4.2 Identification of alternative splicing events affecting transcription factor's binding domains

Using API scripts of Ensembl BioMart the alternative splicing events for each TFs were retrieved. Also, using Ensembl API, we retrieved the CDS coordinates (on the genome) and the coding exons in between. We finally linked this information with the domain coordinates. This allowed the selection of the alternative splicing events involving only DBDs.

5.4.3 DNA-binding multiple alignment and transcription factor subfamilies identification.

We downloaded the protein sequences (FASTA file) of all the proteins encoded by the transcription factor genes, using Ensembl BioMart []. We implemented a script Perl, which use BioPerl modules, to extract the exact sequences of all the DBDs. We then used Prank alignment [], with the default parameters, to perform a multiple alignment of those sequences. The resulted alignment was used as input to reconstruct a phylogenetic tree for each families. To this aim PhyML software [] was used.
MEME suite [] was used (with default parameters) to discover motifs among DBD belonging to the same class.

Chapter 6

Summary and conclusions

In this thesis, we investigated several aspects of genome-complexity in terms of gene amplification. In particular, we focused our analyses on two biological mechanisms important for the gain of novel functionalities: gene duplication and alternative splicing. After introducing the general topic of genome-complexity and the major mechanisms to reach it (Chapter 1), we performed a genome-wide analysis of gene *A.thaliana* in terms of paralogs (Chapter 2) and singleton genes (Chapter 3). This was followed by an analysis of *Arabidopsis* transcription factor genes (Chapter 4), mainly based on the results obtained in the previous two chapters. To investigate on the impact of alternative splicing on genome-complexity, we studied the extent to which human transcription factor genes are regulated by this complex mechanism (Chapter 5). In Chapter 5 we presented also preliminary results on the comparison between gene duplication and alternative splicing in *A.thaliana* and human, focusing particular attention on the different usage of the two mechanisms in the two considered species.

6.1 Genome duplication and gene annotation: an example for a reference plant species.

Arabidopsis thaliana is the reference in plant genomics since its complete genome sequence was the first one to be made available in 2000. As a reference model the *Arabidopsis* genome should be fully reliable and safely annotated; moreover its organization should be well understood in terms of evolutionary mechanisms that gave rise to the actual genome structure. However the presence of widespread intragenome duplications, together with the loss of many gene copies, associated to

possible ancient and recent polyploidization events, really complicates the interpretation of the factors contributing to the genome shaping, thus limiting also the role of this genome as a reference in plant comparative genomics.

To further exploit the available information, and with the aim of supporting the genome annotation of other plant and crop species, we investigated the organization of the *Arabidopsis* genome in terms of paralog genes. We identified all the possible pair-wise similarities between genes classifying structurally related ones into networks, with each gene belonging to only one network given the presence of one or more paralogy relationships. We also focused on the identification of single copy genes (singletons), because their presence in a highly duplicated genome is still an intriguing evolutionary issue.

In order to classify duplicated genes first and obtain and analyze single copy gene, we implemented two adequate pipelines. This is not a novelty in plant genomics [Duarte JM, 2010], but the dedicated pipelines we designed represent a step forward since it faces well-known and still not exhaustively discussed issues related to the computational assessments of paralogs and to peculiarities of the specific methods employed [Van de Peer Y., 2004].

The organization of duplicated genes in networks, are a useful tool for the elucidation and the unraveling of the *Arabidopsis thaliana* gene content in terms of pair-wise paralogy relationships. In fact, networks represent a first step to investigate genes and/or gene families' structural relationships and evolution. Furthermore, our results improve the use of *Arabidopsis* as reference plant species, since the understanding of *Arabidopsis* genome organization is mandatory when it is used as model for other plant species analyses.

Moreover, networks analysis highlights several evolutionary and functional issues that represent the basis for further investigations. For example, the identification of a large network containing approximately one fifth of the *Arabidopsis* genome, associated with the presence of a huge number of "two-gene networks", i.e. networks made of only two genes, are two contrasting but important evolutionary evidence, which arise several questions about the mysterious past of this plant species.

6.2 Approaching gene organization in a highly duplicated genome: an example for transcription factors in *A. thaliana*

We applied the results obtained from the genome-wide analysis of *A.thaliana* to study transcription factor gene families and their organization in such a complex and duplicated genome. We first got a new classification of the transcription factor repertoire in *A.thaliana* since the lack of an exhaustive and comprehensive available dataset. This challenge is compounded by the presence of many dedicated databases methods for the identification of genes encoding DNA-binding domains. Through the integration of data from different databases first, and with the domain analysis then, we provide a solid classification of TFs and coregulators in *A.thaliana*.

Our analysis expanded the existing knowledge about transcription factor genes since we studied their families in terms of gene duplication: we identified structural relationships between TFs belonging to different families and/or between transcription factors and coregulators as well as relationships among transcription factors/coregulators and other classes of genes.

Moreover, transcription factors and coregulators share paralogies also with orphan genes, ie. transcription factor genes not annotated yet. The identification of such relationships together with the analysis of the InterPro domains involved in the paralogy, allowed the refinement of the classification of some orphan genes, supporting the *Arabidopsis thaliana* annotation.

6.3 A web-based database for a deep analysis of the *Arabidopsis thaliana* genome

All the results obtained by the analyses on the *Arabidopsis thaliana* genome can be accessed through a user-friendly and intuitive database. Here the distribution of genes among networks, the classification of singleton genes, as well as the results about transcription factor gene families can be investigated in several ways.

This tool allows the scientific community to use our results for different purpose and researches, contributing to the understanding of plant genomics. In particular, this is the first web resource, in our opinion, that gives the opportunity to view genes in the context of network of paralogs. This novel view of paralogy relationships between

genes can be a helpful tool for researchers working in genome annotation and gene family studies. Moreover, our analysis on transcription factor genes can be used as examples to study all the other *Arabidopisis* gene families, their organization and their connectivity within the genome.

6.4 Investigation on the impact of alternative splicing on human transcription factor genes

Alternative splicing of TFs is known to have a big impact on their regulatory function, particularly in higher organisms such as humans. However, apart from individual examples studied using molecular techniques, the extent and complexity of TF-splicing is still poorly understood on a genomic scale.

Here we presented a focusing analysis aimed to explain how alternative splicing can regulate transcription factors acting on their modularity, and in particular on the presence of the DNA-binding domain.

The classification of the transcription factors in terms of number of isoforms was the first result, based on a careful and dedicated manually curated analysis.

The proposed analysis is the first genome-scale analysis in any mammalian organism of how transcription factors are alternatively spliced and the impact this has on gene regulation. The identification of alternative splicing events involved in the TF regulatory activity is a step forward for the understanding of the transcriptional machinery.

Furthermore, our results will provide the basis of many future computational and experimental studies of the dynamics of splicing in different human cell-types and time-course of development. Moreover it can provide a starting point to investigate whether unusual TF isoforms arise in diseases such as cancer.

Finally we presented preliminary results of a comparative analysis between alternative splicing and gene duplication, as two of the major mechanisms for the proteomic complexity. Particularly interesting earlier evidences of a different usage of the two mechanisms in *Arabidopsis* thaliana and in human. When the analysis was repeated using *Arabidopsis* and human transcription factor genes, we still noticed a different behaviour in the two species.

References

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000 Dec 14;408(6814):796-815.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. PubMed PMID: 2231712.

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell. 2000 Jul;12(7):1093-101.

Blanc G, Hokamp K, Wolfe KH. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 2003 Feb;13(2):137-44.

Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16, 1679–1691

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol Biol. 2010 Feb 24;10:61.

Gaut BS. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res. 2001 Jan;11(1):55-66.

Gibson TJ, Spring J. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. 1998 Feb;14(2):46-9; discussion 49-50.

Grant D, Cregan P, Shoemaker RC. Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci U S A. 2000 Apr 11;97(8):4168-73.

He X, Zhang J. Gene complexity and gene duplicability. Curr Biol. 2005 Jun 7;15(11):1016-21.

Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645

Ku HM, Vision T, Liu J, Tanksley SD. Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci U S A. 2000 Aug 1;97(16):9121-6.

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes Science. 2000 Nov 10;290(5494):1151-5.

Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000 Jan;154(1):459-73.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418-426.

Pearson W. Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics. 2004 Feb; Chapter 3:Unit3.9.

Rubin GM, et al. Comparative genomics of the eukaryotes. Science. 2000 Mar 24;287(5461):2204-15.

Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, et al. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell. 1987 Aug 28;50(5):667.

Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb;12(2):85-94.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011 Feb 1;27(3):431-2. Epub 2010 Dec 12.

Terryn N, Heijnen L, De Keyser A, Van Asseldonck M, De Clercq R, Verbakel H, Gielen J, Zabeau M, Villarroel R, Jesse T, Neyt P, Hogers R, Van Den Daele H, Ardiles W, Schueller C, Mayer K, Déhais P, Rombauts S, Van Montagu M, Rouzé P, Vos P. Evidence for an ancient chromosomal duplication in Arabidopsis thaliana by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett. 1999 Feb 26;445(2-3):237-45.

The Arabidopsis Information Resource (TAIR) [ftp://ftp.Arabidopsis.org/home/tair/Genes/TAIR9_genome_release/], on www.Arabidopsis.org, [Jun 15, 2009]

The Arabidopsis Information Resource (TAIR), [ftp://ftp.Arabidopsis.org/home/tair/Genes/TAIR10_genome_release/], on www.Arabidopsis.org, [Nov 17, 2010]

Vandepoele K, Saeys Y, Simillion C, Raes J, Van De Peer Y. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res. 2002 Nov;12(11):1792-801.

Van de Peer Y. Computational approaches to unveiling ancient genome duplications. Nat Rev Genet. 2004 Oct;5(10):752-63. Review.

Van de Peer, Y., Meyer, A. (2005) Large-scale gene and ancient genome duplications. Book chapter in: The Evolution of the Genome, edited by T.R. Gregory. Elsevier, San Diego.

Vision TJ, Brown DG, Tanksley SD. The origins of genomic duplications in Arabidopsis. Science. 2000 Dec 15;290(5499):2114-7.

Wagner A. Selection and gene duplication: a view from the genome. Genome Biol. 2002;3(5) Apr 15. Review.

Wolfe KH. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2001 May;2(5):333-41

Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladek R. TFCat: the curated catalog of mouse and human TFs. Genome Biol.

2009;10(3):R29

Koscielny G, Le Texier V, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, Stanley E, Fallsehr C, Hofmann O, Kull M, Harrington E, Boué S, Eyras E, Plass M, Lopez F, Ritchie W, Moucadel V, Ara T, Pospisil H, Herrmann A, G Reich J, Guigó R, Bork P, Doeberitz MK, Vilo J, Hide W, Apweiler R, Thanaraj TA, Gautheret D. ASTD: The Alternative Splicing and Transcript Diversity database. Genomics. 2009 Mar;93(3):213-20. Epub 2008 Dec 24.

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003 Jan 1;31(1):374-8.

Schaefer U, Schmeier S, Bajic VB. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. Nucleic Acids Res. 2011 Jan;39(Database issue):D106-10